SIAM J. MATH. ANAL.
Vol. 17, No. 1, January 1986

ON THE EXISTENCE OF CONJUGATE POINTS FOR A SECOND ORDER ORDINARY DIFFERENTIAL EQUATION*

ANGELO B. MINGARELLI†

Abstract. In this paper we show that a result of S. W. Hawking and R. Penrose [Proc. Roy. Soc. London Ser. A, 314 (1970), pp. 529–548] on the existence of conjugate points for a real second order linear differential equation is a consequence of a much earlier result of M. Yelchin [5]. Yelchin's original proof is clarified and corrected and his result is extended. As a result, we obtain extensions of the Hawking–Penrose theorem and Tipler's [J. Differential Equations, 30 (1978), pp. 165–174] a complementary result to said theorem.

AMS-MOS subject classifications (1980). Primary 34C10, secondary 83F05

Key words. conjugate points, disconjugacy

1. Introduction. In 1970 S. W. Hawking and R. Penrose ([3, p. 541], Hawking and Ellis [2, p. 98]) proved the following interesting result concerning the existence of conjugate points on $(-\infty, \infty)$ for

$$y'' + q(x)y = 0. \tag{1.1}$$

PROPOSITION 1 (Hawking and Penrose). *Let* q: $(-\infty, \infty) \to [0, \infty)$, q *continuous on* $(-\infty, \infty)$ *and* $q(t_1) > 0$ *for at least one point* $t_1 \in \mathbb{R}$. *Then* (1.1) *is not disconjugate on* $(-\infty, \infty)$.

We recall that (1.1) is said to be (Wintner) *disconjugate* on $(-\infty, \infty)$ provided every nontrivial solution has at most one zero on $(-\infty, \infty)$. It is nondisconjugate or, more simply, not disconjugate on $\mathbb{R}$ otherwise, i.e., there exists at least one nontrivial solution of (1.1) which has at least two zeros in $(-\infty, \infty)$.

The relevance of the study of conjugate points of (1.1) to general relativity has been pointed out in [4]. For example, the Jacobi equation

$$\frac{d^2 Z^\alpha}{dt^2} = R_{\alpha a \beta b} V^a Z^\beta V^b$$

which is defined along a timelike geodesic (see [2]) $\gamma(t)$ on which Z^α is a Jacobi field, t is the proper time along $\gamma(t)$, V^a is a unit tangent vector to $\gamma(t)$ and $R_{a\alpha\beta b}$ is the Riemann tensor, has a solution which vanishes at two points $t_1 < t_2$, $t_i \in I$, if and only if (for example, [4]) (1.1) has a solution $y \not\equiv 0$ vanishing at t_1 and t_2 where

$$q = \frac{1}{3}\left(R_{ab} V^a V^b + 2\sigma^2\right)$$

and R_{ab} is the Ricci tensor and σ^2 some nonnegative function. The existence of such conjugate points for (1.1) with q defined above is related to the incompleteness of timelike geodesics via the Avez–Hawking theorem (see [4] for further details).

*Received by the editors January 17, 1984, and in revised form October 10, 1984. This research was supported in part by the Natural Sciences and Engineering Research Council of Canada under grant UO 167.

† Department of Mathematics, University of Ottawa, Ottawa, Ontario, Canada K1N 9B4.

An elementary proof of Proposition 1 is given by F. J. Tipler [4, Thm. 1] who complements the Hawking–Penrose result with

PROPOSITION 2 (Tipler). *Let* q: $(-\infty,\infty)\to(-\infty,\infty)$, q *continuous on* $(-\infty,\infty)$ *and*

$$\int_{-\infty}^{\infty} q(s)\,ds>0, \tag{1.2}$$

(*where the integral is to be understood in the sense*

$$\liminf_{t_1\to-\infty}\left(\liminf_{t_2\to+\infty}\int_{t_1}^{t_2} q(s)\,ds\right)>0$$

whenever (1.2) *fails to converge*). *Then* (1.1) *is not disconjugate on* $(-\infty,\infty)$.

Tipler's results are reconsidered, in part, in a subsequent paper by C. D. Ahlbrandt, D. B. Hinton and R. T. Lewis [1] who present a *finite* interval analogue of Proposition 1 for the weighted equation

$$(y'(x)/r(x))'+q(x)y(x)=0, \qquad -\infty\leqq a<x<b\leqq\infty,$$

which, however, does not allow the case $r(x)=1$, [1, Thm. 3.3].

2. The aim of this paper is to show that, in fact, Propositions 1 and 2 above are consequences of a much earlier result of M. Yelchin [5, Thm. 1] which we call Yelchin's theorem and which we extend in Theorem 3.1. It is to be noted that Yelchin's proof of said theorem causes difficulties as there are many misprints, some of which appear to be errors on the author's part. However his proof can still be saved and the results duly extended to the case when q is merely required to be locally Lebesgue integrable on (a,b), (a finite or infinite interval).

Furthermore, we emphasize that Yelchin's results in [5] actually yield a *necessary and sufficient* condition for (1.1) to be nondisconjugate on (a,b). However, the stated condition is not wholly dependent upon q, as one may expect.

3. In the sequel all integrals are Lebesgue integrals and $AC_{\text{loc}}(a,b)$ stands for the class of all real-valued functions defined on (a,b) and locally absolutely continuous there. The interval (a,b) will be a finite or infinite interval, $-\infty\leqq a<x<b\leqq+\infty$. The (equivalence) class of all functions which are locally Lebesgue integrable on (a,b) will be denoted by $L_{\text{loc}}(a,b)$ and the Lebesgue measure of a measurable set $E\subset(a,b)$ by $\mu(E)$.

Next, let $\mathscr{A}$ denote the collection of all functions ψ: $(a,b)\to\mathbb{R}$ which satisfy (i) and (ii):

$$\text{(i)} \qquad \psi\in AC_{\text{loc}}(a,b), \tag{3.1}$$

$$\text{(ii)} \qquad \lim_{\substack{\beta\to b^-\\ \alpha\to a^+}}\left[\arctan\int_{x_0}^{x}\exp\left\{-2\int_{x_0}^{t}\psi(s)\,ds\right\}dt\right]_{x=\alpha}^{x=\beta}=\pi, \tag{3.2}$$

and $x_0\in(a,b)$ is a fixed (but otherwise arbitrary) point. The main result of this paper is

THEOREM 3.1a). *Let* $q\in L_{\text{loc}}(a,b)$. *Furthermore let* $\psi\in\mathscr{A}$ *satisfy*

$$\psi'(x)+\psi^2(x)+q(x)\geqq 0 \quad a.e.\ on\ (a,b) \tag{3.3}$$

with strict inequality holding on sets E^-, E^+ ($E^-\subset(a,x_0]$, $E^+\subset[x_0,b)$) *with* $\mu(E^{\pm})>0$. *Then* (1.1) *is not disconjugate on* (a,b).

THEOREM 3.1b). *Conversely, let* (1.1) *be nondisconjugate on* (a,b). *Then there exists a function* $\psi \in \mathscr{A}$ *which satisfies* (3.3) *a.e. on* (a,b).

As an immediate consequence we obtain Yelchin's theorems [5, Thms. 1, 2] which are stated there for *continuous* q.

COROLLARY 3.1. *Let* $q \in L_{\text{loc}}(-\infty,\infty)$, q: $(-\infty,\infty)\to[0,\infty)$ *and* $q(x)>0$ *on some measurable set* $E \subset (-\infty,\infty)$ *with* $\mu(E)>0$. *Then* (1.1) *is not disconjugate on* $(-\infty,\infty)$.

Remark 1. The above corollary contains the Hawking–Penrose results. For if q is continuous on $(-\infty,\infty)$, $q(x)\geqq 0$ and $q(t_1)>0$ for some point t_1, then $q(x)>0$ in some finite interval about t_1, by continuity. Such an interval has positive Lebesgue measure and hence Corollary 3.1 applies.

Remark 2. That the Hawking–Penrose result is a direct consequence of Yelchin's theorem is seen by choosing $\psi = 0$ in Theorem 3.1a) and applying a continuity argument to (3.3) in order to obtain strict inequality. (Note that (3.2) is trivially satisfied for this choice of ψ on $(-\infty,\infty)$).

COROLLARY 3.2. *Let* $q \in L_{\text{loc}}(-\infty,\infty)$, q: $(-\infty,\infty)\to(-\infty,\infty)$, *and that* q *satisfies* (3.4) *and* (3.5) *below* (*for some* $x_0 \in \mathbb{R}$)

$$\liminf_{x\to+\infty} \int_{x_0}^{x} q(s)\,ds > 0, \tag{3.4}$$

$$\liminf_{x\to-\infty} \int_{x}^{x_0} q(s)\,ds > 0. \tag{3.5}$$

Then (1.1) *is not disconjugate on* $(-\infty,\infty)$.

Remark 3. Note that Corollary 3.2 includes Proposition 2, cf. [4].

Remark 4. That Tipler's theorem (Proposition 2) is a direct consequence of Yelchin's theorem can be realized by choosing $\psi(x) = -\int_{x_0}^{x} q(s)\,ds$. This choice of ψ then satisfies all the hypotheses of Theorem 3.1a) and so Tipler's theorem follows (see the proof of Corollary 3.2 in §4 of this paper for details).

As a final application we give

PROPOSITION 3. *Let* $q \in L_{\text{loc}}(-\infty,\infty)$ *and assume that* $\int_{x_0}^{x_1} q(s)\,ds \neq 0$ *for some* x_0, x_1. *If*

$$-\infty < \liminf_{|x|\to+\infty} \int_{x_0}^{x}\int_{x_0}^{s} q(t)\,dt\,ds \leqq \liminf_{|x|\to+\infty} \int_{x_0}^{x}\int_{x_0}^{s} q(t)\,dt\,ds < +\infty. \tag{3.6}$$

then

$$y'' + \lambda q(t)\,y = 0 \tag{3.7}$$

is not disconjugate on $(-\infty,\infty)$ *for each real* $\lambda \neq 0$.

4. Proofs.

Proof of Theorem 3.1a). Let u be the solution of (1.1) satisfying the initial conditions $u(x_0)=1$, $u'(x_0)=\psi(x_0)$. Since $\psi \in \mathscr{A}$ the function

$$\Psi(x,\alpha) \equiv \psi'(x)+\psi^2(x)+\alpha\big(-q(x)-\psi'(x)-\psi^2(x)\big)$$

where $\alpha \in [0,1]$ is in $L_{\text{loc}}(a,b)$, as a function of x. Thus the initial value problem

$$u' = \Psi(x,\alpha)\,u, \tag{4.1}$$

$$u(x_0,\alpha)=1, \qquad u'(x_0,\alpha)=\psi(x_0), \tag{4.2}$$

where the prime always denotes differentiation with respect to x, admits a unique solution $u=u(x,\alpha)\in AC_{\text{loc}}(a,b)$ for which $u'(x,\alpha)\in AC_{\text{loc}}(a,b)$ and $u(x,\alpha)$ satisfies (4.1) a.e. for $a<x<b$. Note that for $\alpha=1$, (4.1) reduces to (1.1) while for $\alpha=0$, the solution $u(x,0)=\exp\{\int_{x_0}^{x}\psi(s)\,ds\}$.

We now use the Prüfer-type transformation in (4.1): $u=\rho\cos\phi$, $\phi(x_0)=0$. Let v be a linearly independent solution of (4.1) given by the solution of the differential equation $u'v-uv'=1$, $v(x_0)=0$. Then it is readily verified that ϕ, $\phi'\in AC_{\text{loc}}(a,b)$ and ϕ satisfies

$$\phi'=\rho^{-2},\quad \rho^2=u^2+v^2, \tag{4.3}$$

while ρ, $\rho'\in AC_{\text{loc}}(a,b)$ satisfies the equation

$$\rho''\rho^{-1}-\rho^{-4}=\Psi(x,\alpha) \tag{4.4}$$

a.e. on (a,b), (in contrast with [5, Eq. 8]). Now (4.3) implies that $\phi(\cdot,\alpha)$ is increasing for $x\in(a,b)$ with $\phi(x,\alpha)<0$ ($\phi(x,\alpha)>0$) for $a<x<x_0$ ($x_0<x<b$), for each α, $0\leqq\alpha\leqq 1$. Now $\phi(x,1)$ is the phase of (1.1) whereas $\phi(x,0)$ is the phase of (4.1)–(4.2) with $\alpha=0$. In fact $\phi(x,0)$ is given explicitly by

$$\phi(x,0)=\arctan\left\{\int_{x_0}^{x}\exp\left(-2\int_{x_0}^{t}\phi(s)\,ds\right)dt\right\}. \tag{4.5}$$

In the following, ρ_α will denote $\partial\rho/\partial\alpha$ while $'=\partial/\partial x$ as usual. Now because of the smoothess of Ψ as a function of α,u and v will enjoy the same property and thus ρ, as given in (4.3) is AC $[0,1]$, as a function of α for each x. So differentiating (4.4) with respect to α one finds, after a straightforward but lengthy calculation, that

$$\rho_\alpha''-\left(\rho''\rho^{-1}-4\rho^{-4}\right)\rho_\alpha=-\rho\left(q+\psi'+\psi^2\right), \tag{4.6}$$

i.e., ρ_α satisfies the second order linear differential equation (4.6) (in x) a.e. on (a,b).

Now a particular solution of (4.6) subject to the initial conditions $\rho_\alpha(x_0,\alpha)=\rho'_\alpha(x_0,\alpha)=0$ is given by

$$\rho_\alpha(x,\alpha)=-\rho(x,\alpha)\frac{\Phi(x,\alpha)}{2} \tag{4.7}$$

(since $\rho(x_0,\alpha)=1$, $\rho'(x_0,\alpha)=\psi(x_0)$) where

$$\Phi(x,\alpha)\equiv\int_{x_0}^{x}\rho^2(s,\alpha)\left[\psi'(s)+\psi^2(s)+q(s)\right]\sin 2\left[\phi(x,\alpha)-\phi(s,\alpha)\right]ds, \tag{4.8}$$

(in contrast with [5, Eqs. (11), (12)]). Note that (4.7) can be solved so as to yield information regarding $\rho(x,1)$ and $\rho(x,0)$ and so on $\phi'(x,1)$ and $\phi'(x,0)$ because of (4.3). Thus, integrating (4.7) with respect to α over $[0,1]$ for a fixed x, one easily finds

$$\left[\phi(x,1)-\phi(x,0)\right]'=\frac{\exp\left[\int_0^1\Phi(x,\alpha)\,d\alpha\right]-1}{\rho^2(x,0)}. \tag{4.9}$$

We now proceed with the proof. To this end assume the contrary, i.e., that (1.1) is disconjugate on (a,b). Then (3.3) and the Sturm comparison theorem imply that, for each value of α in $[0,1]$, there do not exist points $x_1(\alpha)$, $x_2(\alpha)$, $(a<x_1(\alpha)<x_0<x_2(\alpha)<b)$ which are zeros of $u(x,\alpha)$. Hence, by the definition of ϕ for each $x\in(a,b)$ and for each $\alpha\in[0,1]$, there holds

$$|\phi(x,\alpha)|<\frac{\pi}{2}. \tag{4.10}$$

Hence the increasing nature of ϕ and (4.10) now implies that $\sin 2[\phi(x,\alpha)-\phi(s,\alpha)]>0$ for each $x_0<s<x$. This, along with (3.3), implies that $\Phi(x,\alpha)\geqq 0$ for all $x\geqq x_0$ (with a similar argument holding if $x\leqq x_0$). Now for $x\in E^+$ such that $\mu(E^+\cap[x_0,x])>0$ (or $x\in E^-$), $\Phi(x,\alpha)>0$, as the inequality in (3.3) is strict for such x. Therefore, (4.9) implies that $\Phi(x,1)-\Phi(x,0)$ is nondecreasing for $x\in(a,b)$. Since $\mu(E^+)>0$, there exists a point $x^+\in E^+$ such that $\mu(E^+\setminus\{(x_0,x^+)\cap E^+\})>0$, and $\mu(E^+\cap(x,x^+))>0$ (by a simple measure-theoretic argument). If we now integrate (4.9) over $[x^+,x]$, where $x>x^+>x_0$, we obtain

$$\phi(x,1)>\phi(x,0)+\lambda^2 \tag{4.11}$$

where $\lambda^2\equiv\phi(x^+,1)-\phi(x^+,0)$, as the right side of (4.9) is positive on a set of positive Lebesgue measure lying to the right of x^+. We can now choose $x^-\in E^-$ similarly. Then for $x<x^-<x_0$,

$$\phi(x,1)<\phi(x,0)-\mu^2 \tag{4.12}$$

where $\mu^2=\phi(x^-,0)-\phi(x^-,1)$. Since ϕ is monotone we get

$$\lim_{x\to b^-}\phi(x,1)\geqq\lim_{x\to b^-}\phi(x,0)+\lambda^2 \tag{4.13}$$

and

$$\lim_{x\to a^+}\phi(x,1)\leqq\lim_{x\to a^+}\phi(x,0)-\mu^2. \tag{4.14}$$

Combining (4.13)–(4.14) we find (because of (3.2))

$$\lim_{x\to b^-}[\phi(x,1)]-\lim_{x\to a^+}[\phi(x,1)]\geqq\pi+\lambda^2+\mu^2.$$

However this contradicts (4.10) and this completes the proof.

Proof of Theorem 3.1b). This is identical to that corresponding to [5, Thm. 2] and so it is omitted.

Proof of Corollary 3.1. Let $\psi\equiv 0$ on $(-\infty,\infty)$. Then $\psi\in\mathscr{A}$ and clearly (3.3) is satisfied since $q(x)\geqq 0$ a.e. by hypothesis. For E bounded $\mu(E)>0$, there exists a point $x_0\in E$ such that $\mu((-\infty,x_0)\cap E)>0$ and $\mu(E\cap(x_0,\infty))>0$ (again by a straightforward measure-theoretic argument). Thus let $E^-\equiv(-\infty,x_0)\cap E$ and $E^+\equiv E\cap(x_0,\infty)$. Then $q(x)>0$ a.e. on $E^-\cup E^+$. (If E is unbounded and $\mu(E)>0$, there is a bounded subset F of E with $\mu(F)>0$. We can then apply the above argument to F.) Thus strict inequality in (3.3) holds on these sets and so Theorem 3.1a) implies that (1.1) is not disconjugate.

Proof of Corollary 3.2. Define ψ by $\psi(x)\equiv-\int_{x_0}^x q(s)\,ds$. Then $\psi\in AC_{\text{loc}}(-\infty,\infty)$ and (3.4) implies that $\psi(x)\leqq -c$ where $c>0$, provided $x\geqq x_1\geqq x_0$. Thus it is immediate that (3.3) is satisfied, since $\psi^2(x)>0$ on some interval about x_1. We now show that ψ satisfies (3.2). Note that

$$\exp\left\{-2\int_{x_0}^x\psi(s)\,ds\right\}=\exp(-2c_1)\exp\left(-2\int_{x_1}^x\psi(s)\,ds\right)$$

where $c_1=\int_{x_0}^{x_1}\psi(s)\,ds$. But $\psi(x)\leqq -c$, for $x\geqq x_1$ implies that

$$\exp\left\{-2\int_{x_0}^x\psi(s)\,ds\right\}\geqq\exp(-2c_1)\exp(2c(x-x_1)). \tag{4.15}$$

Hence

$$\lim_{x\to+\infty}\int_{x_1}^{x}\exp\left\{-2\int_{x_0}^{t}\psi(s)\,ds\right\}dt=+\infty \tag{4.16}$$

and so the same is true if x_1 is replaced by x_0. If $x\leqq x_0$, note that for $x\leqq x_2\leqq x_0$ (3.5) implies that $\int_x^{x_0} q(s)\,ds\geqq c_2>0$. Thus $\psi(x)\geqq c_2$ for all $x\leqq x_2$. As before we can easily derive that

$$\exp\left\{-2\int_{x_0}^{x}\psi(s)\,ds\right\}\geqq c_3\exp\left(2c_2(x_2-x)\right) \tag{4.17}$$

for all $x\leqq x_2$ where $c_3=\exp(-2\int_{x_0}^{x_2}\psi(s)\,ds)$. Now (4.17) implies that

$$\lim_{x\to-\infty}\int_{x_2}^{x}\exp\left(-2\int_{x_0}^{t}\psi(s)\,ds\right)dt=+\infty. \tag{4.18}$$

Combining (4.16) and (4.18) we obtain (3.2) as required. Thus $\psi\in\mathscr{A}$ and so Theorem 3.1a) implies that (1.1) is not disconjugate on $(-\infty,\infty)$.

Proof of Proposition 3. In order to show this we note that if q is such that for $\psi(x)\equiv\int_{x_0}^{x}-q(s)\,ds$, (3.2) holds, then (1.1) is not disconjugate on $(-\infty,\infty)$. Now $\psi(x_1)\neq 0$ by hypothesis and so (3.3) holds (with strict inequality around the point x_1). Again (3.1) is satisfied. Thus it suffices to show that

$$\lim_{x\to+\infty}\int_{x_0}^{x}\exp\left\{2\int_{x_0}^{x}\int_{x_0}^{t}q(s)\,ds\,dt\right\}dx=+\infty \tag{4.19}$$

with an analogous result for $x\to-\infty$. However (3.6) implies that

$$-M\leqq\int_{x_0}^{x}\int_{x_0}^{t}q(s)\,ds\,dt\leqq M \tag{4.20}$$

provided $|x|\geqq X_0$, where $M>0$. Applying the estimate (4.20) to the left side of (4.19) we indeed obtain (4.19). A similar result holds in the other case. Thus (3.6) implies that (1.1) is not disconjugate on $(-\infty,\infty)$. Now replacing q by λq as in (3.7) we still have $\int_{x_0}^{x_1}(\lambda q)(s)\,ds\neq 0$, provided $\lambda\neq 0$, and (3.6) clearly holds for each λ. Thus (3.7) is not disconjugate on $(-\infty,\infty)$ for each $\lambda\neq 0$ which is what we wished to show.

Note. Condition (3.6) excludes the case when q is of constant sign a.e. on $(-\infty,\infty)$ since, in the latter case, at least one of the quantities in (3.6) is infinite.

Hence (3.6) implies that q must be positive a.e. on a set of positive Lebesgue measure and negative a.e. on a (possibly different) set of positive Lebesgue measure.

REFERENCES

[1] C. D. AHLBRANDT, D. B. HINTON AND R. T. LEWIS, *The effect of variable change on oscillation and disconjugacy criteria with applications to spectral theory and asymptotic theory*, J. Math. Anal. Appl., 81 (1981), pp. 234–277.

[2] S. W. HAWKING AND G. F. R. ELLIS, *The Large-Scale Structure of Space-Time*, Cambridge Univ. Press, Cambridge, 1973.

[3] S. W. HAWKING AND R. PENROSE, *The singularities of gravitational collapse and cosmology*, Proc. Roy. Soc. London Ser. A, 314 (1970), pp. 529–548.

[4] F. J. TIPLER, *General relativity and conjugate ordinary differential equations*, J. Differential Equaticns, 30 (1978), pp. 165–174.

[5] M. YELCHIN, *Sur les conditions pour qu'une solution d'un système linéaire du second ordre possède deux zéros*, Comptes Rendus (Doklady) Acad. Sci.URSS, 51 (1946), pp. 573–576.

SIAM J. MATH. ANAL.
Vol. 17, No. 1, January 1986

GLOBAL SIMPLIFICATION OF A SINGULARLY PERTURBED ALMOST DIAGONAL SYSTEM*

HARRY GINGOLD† AND PO-FANG HSIEH‡

Abstract. Given a singularly perturbed differential system $\varepsilon^h Y' = [D(x,\varepsilon)+\varepsilon^\theta R(x,\varepsilon)]Y$, where D is a diagonal matrix, $x\in J=(a,b)$, $\varepsilon\in S_c=(0,c]$, and h and θ are positive numbers, this paper studies the conditions such that this system can be globally simplified by $Y=(I+P(x,\varepsilon))Z$ into $\varepsilon^h Z' = D(x,\varepsilon)Z$, valid in $J\times S_{c_1}$, $(0<c_1\leqq c)$. The method and results of Gingold [SIAM J. Math. Anal., 9 (1978), pp. 1076–1802] are used in this study. These results are also true for complex x and ε in certain domains. If $\theta=h$, the results hold also even when J contains turning points of the equations.

AMS-MOS subject classifications (1980). Primary 34E15; secondary 34E20

Key words. global simplification, singular perturbation, almost diagonal system, turning point

1. Introduction.

Consider an n-dimensional linear ordinary differential system

$$(E_1) \qquad \varepsilon^h Y' = \left[D(x,\varepsilon)+\varepsilon^\theta(x,\varepsilon)\right]Y, \qquad ' = \frac{d}{dx},$$

where x is a real or complex variable with $x\in J=(a,b)$, (J may be infinite), or $x\in G$, a simply connected domain in the complex plane, ε is a real or complex parameter with $\varepsilon\in S_c=(0,c]$ or $S_{\alpha c}=\{\varepsilon\,|\,|\arg\varepsilon|<\alpha,\ 0<|\varepsilon|<c\}$ and θ and h are positive numbers. Here $D(x,\varepsilon)$ and $R(x,\varepsilon)$ are n by n matrices and

$$(1.1) \qquad D(x,\varepsilon)=\operatorname{diag}\{d_1(x,\varepsilon),\cdots,d_n(x,\varepsilon)\}.$$

Let

$$(1.2) \qquad R(x,\varepsilon)=\left(r_{jk}(x,\varepsilon)\right), \qquad j,k=1,2,\cdots,n.$$

Without loss of generality, we can assume

$$(1.3) \qquad r_{jj}(x,\varepsilon)\equiv 0, \qquad j=1,2,\cdots,n.$$

Otherwise, r_{jj} can be combined in d_j.

DEFINITION. The system (E_1) is said to be a globally almost diagonal system (G.A.D.S.) in $\bar{J}$ [or in $\bar{G}$] if there exists an n by n matrix $P(x,\varepsilon)$ in the class of $C^1(\bar{J}\times S_{c_1})$, [or $C^1(\bar{G}\times S_{\alpha c_1})$], $(0<c_1\leqq c)$, such that for a suitable norm

$$(1.4) \qquad \lim\|P\|=0 \quad \text{as } \varepsilon\to 0 \text{ in } S_{c_1}\left[\text{or in } S_{\alpha c_1}\right],$$

uniformly for $x\in\bar{J}$ [or $x\in\bar{G}$] and the following relations hold:

$$(1.5) \qquad Y=(I+P(x,\varepsilon))Z \qquad (I\text{: identity matrix}),$$

$$(E_2) \qquad \varepsilon^h Z' = D(x,\varepsilon)Z.$$

* Received by the editors November 10, 1983, and in revised form April 9, 1984.

† Department of Mathematics, West Virginia University, Morgantown, West Virginia 26506. The work of this author is partially supported by a Senate Research Grant, West Virginia University.

‡ Department of Mathematics, Western Michigan University, Kalamazoo, Michigan 49008. The work of this author is partially supported by a Faculty Research Fellowship, Western Michigan University.

Namely, (E_1) has the fundamental matrix in the following form:

$$Y=(I+P(x,\varepsilon))\exp\left\{\varepsilon^{-h}\int^{x}D(t,\varepsilon)\,dt\right\}. \tag{1.6}$$

The existence of $P(x,\varepsilon)$ and validity of (1.4) on $\bar{J}$ [or on $\bar{G}$] justify the "globality" in contrast to "local" results obtained in a subset of J [or of G].

We shall investigate in this paper the sufficient conditions that (E_1) is a G.A.D.S. for both x in J and x in G.

It is important to identify a *final* stage of "asymptotic decomposition" of the given system (E_1) *as soon as possible*. By knowing this, we can avoid laborious linear transformations and laborious calculations of eigenvalues of n by n matrix functions. This is one of our aims in defining G.A.D.S.

2. Main theorems—real case. For $x,t\in\bar{J}$, $\varepsilon\in S_c$, let

$$D_{jk}(x,t,\varepsilon)=\mathrm{Re}\left\{\varepsilon^{-h}\int_{t}^{x}\left(d_j(s,\varepsilon)-d_k(s,\varepsilon)\right)ds\right\},\quad j,k=1,2,\cdots,n\quad(j\neq k). \tag{2.1}$$

The following assumption will be used.

Assumption 1. Assume that $d_j(x,\varepsilon)\in C(J\times S_c)$. For each pair (j,k), $(j\neq k;\ j,k=1,2,\cdots,n)$, there exist two fixed numbers m_{jk} and $\hat{m}_{jk}$ such that

$$\hat{m}_{jk}\leqq D_{jk}(x,t,\varepsilon)\leqq m_{jk} \tag{2.2}$$

for all $x,t\in\bar{J}$, $\varepsilon\in S_c$ or if $D_{jk}(x,t,\varepsilon)$ is unbounded, then for all $x,t\in J$, $\varepsilon\in S_c$, with either $x<t$ or $x>t$,

$$D_{jk}(x,t,\varepsilon)\leqq m_{jk}\quad\text{or}\quad\hat{m}_{jk}\leqq D_{jk}(x,t,\varepsilon). \tag{2.3}$$

For a suitable norm, let

$$r(\varepsilon)=\int_{a}^{b}\|R(t,\varepsilon)\|dt. \tag{2.4}$$

We have the following theorems for real x.

THEOREM 1. *Assume that*: (i) $D(x,\varepsilon)$ *and* $R(x,\varepsilon)$ *are in the class* $C(\bar{J}\times S_c)$, (ii) *Assumption* 1 *holds*, (iii) $\theta=h$, *and* (iv) $r(\varepsilon)=o(1)$ *as* $\varepsilon\to 0^{+}$. *Then*, (E_1) *is a* G.A.D.S., *where* $\|P(x,\varepsilon)\|=O(r(\varepsilon))$ *uniformly on* $\bar{J}$ *as* $\varepsilon\to 0^{+}$.

THEOREM 2. *Assume that*: (i) $D(x,\varepsilon)$ *and* $R(x,\varepsilon)$ *are in the class* $C^1(\bar{J}\times\bar{S}_c)$, (ii) *Assumption* 1 *holds*, (iii) $\theta>h/2$, (iv) *there is a fixed number* μ *such that*

$$\left|d_j(x,\varepsilon)-d_k(x,\varepsilon)\right|\geqq\mu>0,\qquad j,k=1,2,\cdots,n,\quad j\neq k \tag{2.5}$$

for all $x\in J$ *and* $\varepsilon\in S_c$. *Then* (E_1) *is a* G.A.D.S. *and* $\|P(x,\varepsilon)\|=O(\varepsilon^{\sigma})$ *uniformly on* $\bar{J}$ *as* $\varepsilon\to 0^{+}$, *where* $\sigma=\min\{\theta,2\theta-h\}$.

We shall prove Theorem 1 in §3 and use Theorem 1 to prove Theorem 2 in §7.

Combine the idea of G.A.D.S. with those theorems, we have the next theorem.

THEOREM 3. *Under the conditions of Theorem* 1 *or those of Theorem* 2, *the differential system* (E_1) *has the fundamental matrix with asymptotic expansion* (1.6) *uniformly on* $\bar{J}$.

Analogous theorems with complex x will be given as Theorem 4 and Theorem 5 in §8.

Theorem 1 and Theorem 4 include the case that (E_1) has certain types of turning points in J or in G. Extension of the method used in proving these theorems can be applied to the results of Wasow [15] and Lee [10]. Also extensions of the method and results presented in these theorems can be applied to improve those of Devinatz [2], Levinson [11], Harris and Lutz [6], and Hartman and Wintner [8]. These will be discussed in forthcoming papers.

The question of finding the "leading term" of the coefficient of (E_1) has been discussed by many over the years (see Hsieh [8]). The theorems in the paper indicate that $D(x,\varepsilon)$ is the leading term of the coefficients of (E_1) if $\theta > h/2$ and appropriate additional conditions are satisfied.

3. Proof of Theorem 1. From the equations (E_1), (1.5), (E_2) and $\theta = h$, we know that $P(x,\varepsilon)$ satisfies the following equation:

$$\varepsilon^h P' = D(x,\varepsilon)P - PD(x,\varepsilon) + \varepsilon^h R(x,\varepsilon)P + \varepsilon^h R(x,\varepsilon). \tag{3.1}$$

Let

$$E(x,t,\varepsilon) = \exp\left\{\varepsilon^{-h}\int_t^x D(s,\varepsilon)\,ds\right\} \tag{3.2}$$

and L be the integral operator

$$LP = \int^x E(x,t,\varepsilon)R(t,\varepsilon)P(t,\varepsilon)E^{-1}(x,t,\varepsilon)\,dt \tag{3.3}$$

with lower limits to be specified in the sequel. Then, by a well-known lemma (e.g. see Wasow [14, p. 169]), the solution P of (3.1) is given by

$$P(x,\varepsilon) = P_0 + LP, \qquad P_0 = LI. \tag{3.4}$$

Let

$$P = (p_{jk}), \quad RP = ((RP)_{jk}), \quad LP = ((LP)_{jk}), \qquad j,k = 1,2,\cdots,n. \tag{3.5}$$

Then

$$(LP)_{jk} = \int_{\beta_{jk}}^x (RP)_{jk} \exp\left\{\varepsilon^{-h}\int_t^x \left(d_j(s,\varepsilon) - d_k(s,\varepsilon)\right) ds\right\} dt, \tag{3.6}$$

where β_{jk} are chosen to be either a or b so that for t between β_{jk} and x, the exponential factor of the integrand remains bounded. Let

$$m = \max_{j,k}\left\{|m_{jk}|, |\hat{m}_{jk}|\right\} \tag{3.7}$$

and the norm is chosen that

$$|r_{jk}| \leqq \|R\| \quad \text{and} \quad \|R\| \leqq \left\|\left(|r_{jk}|\right)\right\|, \tag{3.8}$$

also, denote

$$|||P||| = \sup_J \|P\|. \tag{3.9}$$

Then, we have the estimate

$$(3.10)\qquad |(LP)_{jk}| \leqq \left| \int_{\beta_{jk}}^{x} (RP)_{jk}(\exp m)\, dt \right| \leqq e^m \sum_{l=1}^{n} \left| \int_{\beta_{jk}}^{x} |r_{jl}(t,\varepsilon)||p_{lk}(t,\varepsilon)|dt \right|$$

$$\leqq ne^m \left| \int_{\beta_{jk}}^{x} \|R\| \left(\max_{l,k} |p_{lk}| \right) dt \right| \leqq ne^m |||P||| \left| \int_{\beta_{jk}}^{x} \|R\| dt \right|.$$

Let

$$(3.11)\qquad K = ne^m.$$

We have

$$(3.12)\qquad \|LP\| \leqq K|||P||| \int_a^b \|R(t,\varepsilon)\| dt.$$

Similarly, we have

$$(3.13)\qquad \|LI\| \leqq K\|I\| \int_a^b \|R(t,\varepsilon)\| dt.$$

Thus, from (3.4), and (2.3) we have

$$(3.14)\qquad |||P||| \leqq K(\|I\| + |||P|||) r(\varepsilon).$$

By Theorem 1 (iv), if c_1 is chosen small enough, (3.4) defines a contraction mapping for $x \in \bar{J}$, $\varepsilon \in S_{c_1}$. Furthermore, we have

$$(3.15)\qquad |||P||| \leqq \frac{K\|I\| r(\varepsilon)}{1 - Kr(\varepsilon)}$$

and Theorem 1 is proved.

4. Fundamental lemmas. In order to prove Theorem 2, we need to establish the following fundamental lemmas.

LEMMA 1. *Let Z and P_{k0} be n by n matrices with $Z = (z_{jk})$, $j,k = 1,2,\cdots,n$ and P_{k0} is the matrix with one at its (k,k) entry and zero everywhere else. Then, Z commutes with P_{k0} if, and only if, that*

$$(4.1)\qquad z_{jk} = z_{kr} = 0 \quad \text{for } j, r \neq k.$$

LEMMA 2. *Assume that $D(x,\varepsilon)$ and $R(x,\varepsilon)$ satisfy* (1.1)–(1.3) *and* (i) *and* (iv) *of Theorem* 2. *Then*: (1) *the matrix $D(x,\varepsilon)+\eta R(x,\varepsilon)$ has distinct eigenvalues $\{\tilde{d}_1(x,\varepsilon,\eta), \tilde{d}_2(x,\varepsilon,\eta),\cdots,\tilde{d}_n(x,\varepsilon,\eta)\}$ for $x \in \bar{J}$, $\varepsilon \in S_c$, $|\eta| \leqq \eta_0$ with a suitable positive constant η_0*; (2) *the characteristic polynomial of $D(x,\varepsilon)+\eta R(x,\varepsilon)$ has the form*

$$(4.2)\qquad \tilde{p}_n(\lambda; x,\varepsilon,\eta) \equiv \det[D + \eta R - \lambda I] = p_n(\lambda; x,\varepsilon) + \eta^2 p_{n-2}(\lambda; x,\varepsilon,\eta),$$

where

$$(4.3)\qquad p_n(\lambda; x,\varepsilon) = \prod_{j=1}^{n} \left(d_j(x,\varepsilon) - \lambda \right),$$

$p_{n-2}(\lambda; x,\varepsilon,\eta)$ *is a polynomial in λ of degree $n-2$; and* (3) $\tilde{d}_j(x,\varepsilon,\eta)$ *satisfy the following relations*;

$$(4.4)\qquad \tilde{d}_j(x,\varepsilon,\eta) = d_j(x,\varepsilon) + O(\eta^2), \qquad j = 1,2,\cdots,n$$

for $x \in \bar{J}$, $\varepsilon \in \bar{S}_c$, $|\eta| \leqq \eta_1$.

LEMMA 3. *Given an n by n matrix* $D(x,\varepsilon)+\eta R(x,\varepsilon)$, *where D and R satisfy conditions* (1.3) *and* (i) *and* (iv) *of Theorem* 2, *and* η *is a complex number. Then, there is an n by n matrix* $Q(x,\varepsilon,\eta)$ *in the class of* $C^1(\bar{J}\times\bar{S}_c\times\{|\eta|\leqq\tilde{\eta}\})$ ($\tilde{\eta}$: *positive constant*) *satisfying*

$$\|Q(x,\varepsilon,\eta)\|\leqq K_1|\eta|, \qquad \|Q'(x,\varepsilon,\eta)\|\leqq K_2|\eta| \tag{4.5}$$

uniformly on $\bar{J}\times\bar{S}_c\times\{|\eta|\leqq\tilde{\eta}\}$ *for suitable positive constants* K_1 *and* K_2 *and*

$$(I+Q)^{-1}(D+\eta R)(I+Q)=\tilde{D}(x,\varepsilon,\eta) \tag{4.6}$$

where

$$\tilde{D}(x,\varepsilon,\eta)=\operatorname{diag}\{\tilde{d}_1(x,\varepsilon,\eta),\tilde{d}_2(x,\varepsilon,\eta),\cdots,\tilde{d}_n(x,\varepsilon,\eta)\}. \tag{4.7}$$

In order to show Lemma 1 subdivide Z according to P_{k0}, namely

$$Z=\begin{pmatrix} Z_{11} & Z_{12} & Z_{13} \\ Z_{21} & z_{kk} & Z_{23} \\ Z_{31} & Z_{32} & Z_{33} \end{pmatrix}, \tag{4.8}$$

where Z_{11} is $k-1$ by $k-1$, Z_{12} is $k-1$ by 1, Z_{13} is $k-1$ by $n-k$, Z_{21} is 1 by $k-1$, Z_{31} is $n-k$ by $k-1$ etc. Then, by equating $ZP_{k0}=P_{k0}Z$, we get that the entries of Z_{12}, Z_{21}, Z_{23}, and Z_{32} are all zero. Thus, Lemma 1 is proved.

5. Proof of Lemma 2. To show Lemma 2, first note that $D+\eta R$ is linear in η, and its eigenvalues are continuous functions of all of its variables (x,ε,η). For $\eta=0$, by (iv), the eigenvalues of D are distinct for $x\in J$ and $\varepsilon\in S_c$, there is a positive constant η_1 such that the eigenvalues $\{\tilde{d}_1(x,\varepsilon,\eta),\tilde{d}_2(x,\varepsilon,\eta),\cdots,\tilde{d}_n(x,\varepsilon,\eta)\}$ are distinct for $x\in J$, $\varepsilon\in S_c$, $|\eta|\leqq\eta_1$.

To show (4.2), by mathematical induction on n, note first that $r_{jj}\equiv 0$, by (1.3). For $n=2$,

$$\tilde{p}_2(\lambda;x,\varepsilon,\eta)=\begin{vmatrix} d_1-\lambda & \eta r_{12} \\ \eta r_{21} & d_2-\lambda \end{vmatrix}=(d_1-\lambda)(d_2-\lambda)-\eta^2 r_{12}r_{21}. \tag{5.1}$$

Thus (4.2) is true for $n=2$. Assume that (4.2) is true for $n=k$. Namely, when D and R are k by k matrices

$$\tilde{p}_k(\lambda;x,\varepsilon,\eta)=p_k(\lambda,x,\varepsilon)+\eta^2 p_{k-2}(\lambda;x,\varepsilon,\eta). \tag{5.2}$$

For $n=k+1$, expand $\tilde{p}_{k+1}(\lambda;x,\varepsilon,\eta)$ with respect to the last column. Then,

$$\begin{aligned}\tilde{p}_{k+1}(\lambda;x,\varepsilon,\eta)&=\det[D+\eta R-\lambda I] \\ &=(d_{k+1}-\lambda)\{p_k(\lambda;x,\lambda)+\eta^2 p_{k-2}(\lambda;x,\varepsilon,\eta)\} \\ &\quad+\eta\sum_{j=1}^{k} r_{j,k+1}\Delta_{j,k+1},\end{aligned} \tag{5.3}$$

where $\Delta_{j,k+1}$ are cofactors of $\eta r_{j,k+1}$. Since $\Delta_{j,k+1}$ is obtained by deleting the jth row and the last column from $\det[D+\eta R-\lambda I]$, it is a polynomial in λ of degree $k-1$ and has η as a common factor. Therefore

$$\eta^2(d_{k+1}-\lambda)p_{k-2}(\lambda;x,\varepsilon,\eta)+\eta\sum_{j=1}^{k} r_{j,k+1}\Delta_{j,k+1}=\eta^2 p_{k-1}(\lambda;x,\varepsilon,\eta), \tag{5.4}$$

where $p_{k-1}(\lambda; x, \varepsilon, \eta)$ is a polynomial in λ of degree $k-1$. Thus, (4.2) is true for all positive integers n.

To show (4.4), following the method given in Gingold [4], consider a new polynomial

$$q_n(\lambda; x, \varepsilon, \eta, \hat{\eta}) = p_n(\lambda; x, \varepsilon) + \hat{\eta} p_{n-2}(\lambda; x, \varepsilon, \eta), \tag{5.5}$$

where $\hat{\eta}$ is a complex number. We can assume, without loss of generality, that $q_n(\lambda; x, \varepsilon, \eta, \hat{\eta}) = 0$ has distinct roots for $x \in \bar{J}$, $\varepsilon \in \bar{S}_c$, $|\eta| \leqq \eta_1$, $|\hat{\eta}| \leqq \eta_2$, where η_2 is a small positive number. Regard these roots as functions of $\hat{\eta}$, and x, ε and η as their parameters. These roots satisfy the initial value problems

$$\frac{d\lambda}{d\hat{\eta}} = \frac{-p_{n-2}(\lambda; x, \varepsilon, \eta)}{\partial q_n / \partial \lambda}, \qquad \lambda(0) = d_j(x, \varepsilon), \qquad j = 1, 2, \cdots, n. \tag{5.6}$$

Then, these n distinct roots are holomorphic in $\hat{\eta}$ and expressible as

$$\lambda_j = d_j(x, \varepsilon) + O(\hat{\eta}), \qquad j = 1, 2, \cdots, n, \tag{5.7}$$

where $O(\hat{\eta})$ is uniform with respect to $x \in \bar{J}$, $\varepsilon \in \bar{S}_c$, $|\eta| \leqq \eta_0$. By letting $\hat{\eta} = \eta^2$, $\eta_0 = \min(\eta_1, \sqrt{\eta_2})$, $\lambda_j = \tilde{d}_j(x, \varepsilon, \eta)$, we have

$$\tilde{d}_j(x, \varepsilon, \eta) = d_j(x, \varepsilon) + O(\eta^2) \tag{4.4}$$

for $x \in J$, $\varepsilon \in S_c$, $|\eta| \leqq \eta_0$.

Thus Lemma 2 is proved.

6. Proof of Lemma 3. To prove Lemma 3, we will recreate the proofs of Gingold [3] carefully to obtain the desired estimates. As proved in Lemma 2, $D(x, \varepsilon) + \eta R(x, \varepsilon)$ has distinct eigenvalues $\{\tilde{d}_1(x, \varepsilon, \eta), \tilde{d}_2(x, \varepsilon, \eta), \cdots, \tilde{d}_n(x, \varepsilon, \eta)\}$ for $x \in J$, $\varepsilon \in S_c$, $|\eta| \leqq \eta_0$. Namely, there exists a fixed number $\hat{\mu}$, such that

$$\inf_{\substack{x \in J,\, \varepsilon \in S_c,\, |\eta| \leqq \eta_0 \\ 1 \leqq j,\, k \leqq n,\, j \neq k}} |\tilde{d}_j(x, \varepsilon, \eta) - \tilde{d}_k(x, \varepsilon, \eta)| \geqq \hat{\mu} > 0. \tag{6.1}$$

Let Γ_k, $k = 1, 2, \cdots, n$, be a set of rectifiable closed Jordan curves in the λ-plane such that Γ_k contains $\lambda = \tilde{d}_k(x, \varepsilon, \eta)$ in its interior and $\lambda = \tilde{d}_j(x, \varepsilon, \eta)$, $(j \neq k)$, in its exterior for all $x \in \bar{J}$, $\varepsilon \in \bar{S}_c$, $|\eta| \leqq \eta_0$. Consider the matrices $P_k(x, \varepsilon, \eta)$ given by

$$P_k(x, \varepsilon, \eta) = \frac{1}{2\pi i} \oint_{\Gamma_k} R_\lambda(x, \varepsilon, \eta)\, d\lambda, \qquad k = 1, 2, \cdots, n, \tag{6.2}$$

where

$$R_\lambda(x, \varepsilon, \eta) = [\lambda I - D - \eta R]^{-1} = R_{\lambda 0}(x, \varepsilon) \sum_{\nu=0}^{\infty} [\eta R_{\lambda 0} R]^\nu \tag{6.3}$$

with

$$R_{\lambda 0}(x, \varepsilon) = [\lambda I - D]^{-1} = \operatorname{diag}\left\{(\lambda - d_1(x, \varepsilon))^{-1}, \cdots, (\lambda - d_n(x, \varepsilon))^{-1}\right\}. \tag{6.4}$$

Then $P_k(x, \varepsilon, \eta)$ is a projection (e.g. see Riesz and Sz-Nagy [12, p. 419]) and in the class of $C^1(\bar{J} \times \bar{S}_c)$ and holomorphic in η for $\{|\eta| \leqq \eta_0\}$ for some positive η_0.

Similar to the differential equation introduced by Kato [9] and used by Coppel [1] and Gingold [3], consider the initial value problems

$$\frac{dW_k}{d\eta} = \left[\frac{dP_k}{d\eta} P_k - P_k \frac{dP_k}{d\eta}\right] W_k, \qquad W_k(x, \varepsilon, 0) = I, \quad k = 1, 2, \cdots, n, \tag{6.5}$$

and denote each of its unique n by n solution by $W_k(x,\varepsilon,\eta)$, $k=1,2,\cdots,n$. By the Cauchy's theorem, the following n by n matrix given by

$$P_{k0}=\frac{1}{2\pi i}\oint_{\Gamma_k} R_{\lambda 0}(x,\varepsilon)\,d\lambda \tag{6.6}$$

has entries all zero except the element at the (k,k) place which is one. Consequently,

$$W_k^{-1}(x,\varepsilon,0)P_{k0}W_k(x,\varepsilon,0)=P_{k0}. \tag{6.7}$$

As shown in Gingold [3], (6.7) implies

$$W_k^{-1}(x,\varepsilon,\eta)P_k(x,\varepsilon,\eta)W_k(x,\varepsilon,\eta)=P_{k0} \tag{6.8}$$

for all η, $|\eta|\leqq\eta_0$. Let

$$Z_k(x,\varepsilon,\eta)=W_k^{-1}(x,\varepsilon,\eta)(D+\eta R)W_k(x,\varepsilon,\eta) \tag{6.9}$$

and

$$Z_k(x,\varepsilon,\eta)=(z_{kij}),\qquad i,j=1,2,\cdots,n. \tag{6.10}$$

Note that $[\lambda I-D-\eta R]^{-1}$ and $D+\eta R$ are commutative, $P_k(x,\varepsilon,\eta)$ and $D+\eta R$ are commutative. By (6.8)

$$\begin{aligned} Z_k(x,\varepsilon,\eta)P_{k0} &= W_k^{-1}(x,\varepsilon,\eta)(D+\eta R)W_k(x,\varepsilon,\eta)P_{k0} \\ &= W_k^{-1}(x,\varepsilon,\eta)(D+\eta R)P_k(x,\varepsilon,\eta)W_k(x,\varepsilon,\eta) \\ &= W_k^{-1}(x,\varepsilon,\eta)P_k(x,\varepsilon,\eta)(D+\eta R)W_k(x,\varepsilon,\eta) \\ &= P_{k0}W_k^{-1}(x,\varepsilon,\eta)(D+\eta R)W_k(x,\varepsilon,\eta) \\ &= P_{k0}Z_k(x,\varepsilon,\eta). \end{aligned} \tag{6.11}$$

Namely, $Z_k(x,\varepsilon,\eta)$ and P_{k0} are commutative. By Lemma 1, we have

$$z_{kik}(x,\varepsilon,\eta)=z_{kkj}(x,\varepsilon,\eta)=0 \quad \text{for } i,j\neq k. \tag{6.12}$$

Also, note that the only nonzero element in $Z_k(x,\varepsilon,\eta)P_{k0}$ is z_{kkk} and P_k is the projection which takes the vector space on which $D+\eta R$ operates onto the subspace spanned by the eigenvector corresponding to $\tilde{d}_k(x,\varepsilon,\eta)$. Also, from (6.9), we have

$$(D+\eta R)W_k(x,\varepsilon,\eta)P_{k0}=W_k(x,\varepsilon,\eta)Z_k(x,\varepsilon,\eta)P_{k0}. \tag{6.13}$$

Therefore,

$$z_{kkk}(x,\varepsilon,\eta)=\tilde{d}_k(x,\varepsilon,\eta) \tag{6.14}$$

and $W_k(x,\varepsilon,\eta)P_{k0}$ has exactly one nonzero column which is the eigenvector of $D+\eta R$ corresponding to $\tilde{d}_k(x,\varepsilon,\eta)$.

Now, denote the coefficient of (6.5) by

$$\frac{dP_k}{d\eta}P_k-P_k\frac{dP_k}{d\eta}=N_k(x,\varepsilon,\eta) \tag{6.15}$$

which is in the class of $C^1(\bar{J}\times\bar{S}_c)$, holomorphic in η for $\{|\eta|\leqq\eta_0\}$. Thus, the solution of (6.5) can be written as

$$W_k(x,\varepsilon,\eta)=I+\int_0^{\eta}N_k(x,\varepsilon,\eta_1)\,d\eta_1 + \int_0^{\eta}N_k(x,\varepsilon,\eta_1)\int_0^{\eta_1}N_k(x,\varepsilon,\eta_2)\,d\eta_2 d\eta_1+\cdots \tag{6.16}$$

which converges uniformly for $|\eta|\leqq\tilde{\eta}_k$, for a certain positive constant $\tilde{\eta}_k$. Let

$$Q_k(x,\varepsilon,\eta)=W_k(x,\varepsilon,\eta)-I. \tag{6.17}$$

Then,

$$Q_k(x,\varepsilon,\eta)=\int_0^{\eta}N_k(x,\varepsilon,\eta_1)\,d\eta_1+\int_0^{\eta}N_k(x,\varepsilon,\eta_1)\int_0^{\eta_1}N_k(x,\varepsilon,\eta_2)\,d\eta_2 d\eta_1+\cdots. \tag{6.18}$$

Since N_k is in the class of $C^1(\bar{J}\times\bar{S}_c)$, and holomorphic in η for $\{|\eta|\leqq\tilde{\eta}_k\}$, there exists a positive constant g_k, independent of (x,ε,η), such that

$$\|N_k(x,\varepsilon,\eta)\|\leqq g_k \quad \text{for } (x,\varepsilon,\eta) \text{ in } \bar{J}\times\bar{S}_c\times\{|\eta|\leqq\tilde{\eta}_k\}. \tag{6.19}$$

Then,

$$\|Q_k(x,\varepsilon,\eta)\|\leqq g_k\left|\int_0^{\eta}d\eta_1\right|+g_k^2\left|\int_0^{\eta}\int_0^{\eta_1}d\eta_2 d\eta_1\right|+\cdots \tag{6.20}$$

$$\leqq e^{g_k|\eta|}-1\to 0 \quad \text{as } |\eta|\to 0.$$

Thus, we can put

$$Q_k(x,\varepsilon,\eta)=\eta G_k(x,\varepsilon,\eta), \tag{6.21}$$

where G_k is in the class of $C^1(\bar{J}\times\bar{S}_c)$ and holomorphic in η for $\{|\eta|\leqq\tilde{\eta}_k\}$.

Put

$$W(x,\varepsilon,\eta)=\sum_{k=1}^{n}W_k(x,\varepsilon,\eta)P_{k0}. \tag{6.22}$$

Then, by (6.9), (6.13) and (6.14), we have

$$\begin{aligned}(D+\eta R)W&=\sum_{k=1}^{n}(D+\eta R)W_k(x,\varepsilon,\eta)P_{k0}\\ &=\sum_{k=1}^{n}W_k(x,\varepsilon,\eta)Z_k(x,\varepsilon,\eta)P_{k0}\\ &=\sum_{k=1}^{n}W_k(x,\varepsilon,\eta)P_{k0}\tilde{D}(x,\varepsilon,\eta)\\ &=W(x,\varepsilon,\eta)\tilde{D}(x,\varepsilon,\eta),\end{aligned} \tag{6.23}$$

where $\tilde{D}$ is given in (4.7), namely

$$\tilde{D}(x,\varepsilon,\eta)=\operatorname{diag}\{\tilde{d}_1(x,\varepsilon,\eta),\cdots,\tilde{d}_n(x,\varepsilon,\eta)\}. \tag{4.7}$$

Put

$$W(x,\varepsilon,\eta)=I+Q(x,\varepsilon,\eta). \tag{6.24}$$

Then, since $\sum_{k=0}^{n} P_{k0} = I$, and by (6.22), we have

$$Q(x,\varepsilon,\eta) = \sum_{k=1}^{n} Q_k(x,\varepsilon,\eta) P_{k0} = \eta \sum_{k=1}^{n} G_k(x,\varepsilon,\eta) P_{k0}, \tag{6.25}$$

$$\begin{aligned} Q'(x,\varepsilon,\eta) &= \sum_{k=1}^{n} Q_k'(x,\varepsilon,\eta) P_{k0} \\ &= \eta \sum_{k=1}^{n} G_k'(x,\varepsilon,\eta) P_{k0}, \qquad \left(' = \frac{d}{dx}\right). \end{aligned} \tag{6.26}$$

Put

$$\tilde{\eta} = \min\{\tilde{\eta}_1, \tilde{\eta}_2, \cdots, \tilde{\eta}_n\}, \tag{6.27}$$

$$K_1 = \sup_{\bar{J}\times\bar{S}_c\times\{|\eta|\leqq\tilde{\eta}\}} \left\| \sum_{k=1}^{n} G_k(x,\varepsilon,\eta) P_{k0} \right\|, \tag{6.28}$$

and

$$K_2 = \sup_{\bar{J}\times\bar{S}_c\times\{|\eta|<\tilde{\eta}\}} \left\| \sum_{k=1}^{n} G_k'(x,\varepsilon,\eta) P_{k0} \right\|. \tag{6.29}$$

Thus, Lemma 3 is proved.

7. Proof of Theorem 2. First of all, since $D(x,\varepsilon)$ is diagonal, by Lemma 3 there exists an n by n matrix $Q(x,\varepsilon)$ in the class of $C^1(\bar{J}\times\bar{S}_c)$ such that

$$(I+Q)^{-1}(D+\varepsilon^{\theta}R)(I+Q) = \tilde{D}(x,\varepsilon), \tag{7.1}$$

where

$$\tilde{D}(x,\varepsilon) = \operatorname{diag}\{\tilde{d}_1(x,\varepsilon), \cdots, \tilde{d}_n(x,\varepsilon)\}. \tag{7.2}$$

Furthermore, we have

$$\|Q(x,\varepsilon)\| = O(\varepsilon^{\theta}), \qquad \|Q'(x,\varepsilon)\| = O(\varepsilon^{\theta}) \tag{7.3}$$

uniformly on $\bar{J}$ as $\varepsilon \to 0^+$.

Now, apply the transformation,

$$Y = (I+Q(x,\varepsilon))V, \tag{7.4}$$

which reduces (E_1) to

$$\begin{aligned} \varepsilon^h V' &= \left[\tilde{D}(x,\varepsilon) - \varepsilon^h (I+Q)^{-1} Q'\right] V \\ &= \left[\hat{D}(x,\varepsilon) + \varepsilon^h \hat{R}(x,\varepsilon)\right] V, \end{aligned} \tag{7.5}$$

where

$$\begin{aligned} \hat{D} &= \tilde{D} - \varepsilon^h \operatorname{diag}(I+Q)^{-1} Q', \\ \hat{R} &= -\left\{(I+Q)^{-1} Q' - \operatorname{diag}(I+Q)^{-1} Q'\right\}. \end{aligned} \tag{7.6}$$

Then, the system (7.5) satisfies the conditions of Theorem 1 with $r(\varepsilon)=O(\varepsilon^{\theta})$. Therefore, there exists an n by n matrix $\tilde{P}(x,\varepsilon)$ in the class of $C^1(\bar{J}\times\bar{S}_{c_2})$, $(0<c_2\leqq c)$, such that

$$Y=(I+Q)(I+\tilde{P})\exp\left\{\varepsilon^{-h}\int^{x}\hat{D}(t,\varepsilon)\,dt\right\}. \tag{7.7}$$

In order to show that (E_1) is a G.A.D.S. following an argument of Gingold [4], we intend to show that

$$\varepsilon^{-h}\int^{x}[\hat{D}(t,\varepsilon)-D(t,\varepsilon)]\,dt=O(\varepsilon^{\sigma}) \tag{7.8}$$

uniformly on $\bar{J}$ as $\varepsilon\to 0^{+}$. In fact, by Lemma 2

$$\tilde{p}_n(\lambda)=\operatorname{Det}(D+\varepsilon^{\theta}R-\lambda I)=p_n(\lambda)+\varepsilon^{2\theta}p_{n-2}(\lambda), \tag{7.9}$$

where

$$p_n(\lambda)=\prod_{j=1}^{n}\left(d_j(x,\varepsilon)-\lambda\right) \tag{7.10}$$

and $p_{n-2}(\lambda)$ is a polynomial of order $n-2$. Since, by (2.5), $p_n(\lambda)$ has distinct zeros d_j, $j=1,2,\cdots,n$, the zeros $\tilde{d}_j$ of $\tilde{p}_n(\lambda)$, as given in (7.2), are also distinct on $\bar{J}\times\bar{S}_{c_3}$ for some c_3 $(0<c_3\leqq c_2)$, and moreover, by suitable indexing of the zeros of $p_n(\lambda)$, we have

$$\tilde{d}_j(x,\varepsilon)=d_j(x,\varepsilon)+O(\varepsilon^{2\theta}),\qquad j=1,2,\cdots,n, \tag{7.11}$$

uniformly on $\bar{J}$ as $\varepsilon\to 0^{+}$. Thus,

$$\begin{aligned}\varepsilon^{-h}\int^{x}&[\hat{D}(t,\varepsilon)-D(t,\varepsilon)]\,dt\\ &=\varepsilon^{-h}\int^{x}[\tilde{D}(t,\varepsilon)-D(t,\varepsilon)]\,dt-\int^{x}\operatorname{diag}(I+Q(t,\varepsilon))^{-1}Q'(t,\varepsilon)\,dt\\ &=O(\varepsilon^{\sigma})\end{aligned} \tag{7.12}$$

uniformly on $\bar{J}$ as $\varepsilon\to 0^{+}$, as the first integral is of $O(\varepsilon^{2\theta-h})$, by (7.11), and the second integral is of $O(\varepsilon^{\theta})$, by (7.3). Let

$$P(x,\varepsilon)=(I+Q)(I+\tilde{P})\left(\exp\left\{\varepsilon^{-h}\int^{x}[\hat{D}(t,\varepsilon)-D(t,\varepsilon)]\,dt\right\}\right)-I, \tag{7.13}$$

thus Theorem 2 is proved.

8. Theorems for the complex variable case. For the case that x is a complex variable, the following assumption will be used.

Assumption 2. Assume that: (1) $D(x,\varepsilon)$ and $R(x,\varepsilon)$ are holomorphic on $\bar{G}\times\bar{S}_{\alpha c}$; (2) there are two fixed points a and b on the boundary of G such that for every $x\in G$, there exists a Jordan curve $\Gamma_x\in G$ connecting x to a and b, and the quantities $D_{jk}(x,t,\varepsilon)$ given by (2.1) are defined for $t\in\Gamma_x$; (3) for each pair j,k $(j\neq k;\ j,k=1,2,\cdots,n)$, there exist two numbers m_{jk} and $\hat{m}_{jk}$ such (2.2) holds for every $x\in G$ and $t\in\Gamma_x$, $\varepsilon\in S_{\alpha c}$, or if $D_{jk}(x,t,\varepsilon)$ is unbounded, then (2.3) holds for all $x\in G$, $t\in\Gamma_x$, $\varepsilon\in S_{\alpha c}$, with t on either side of x.

Let

$$\hat{r}(\varepsilon)=\sup_{\Gamma_x}\int_{\Gamma_x}\|R(t,\varepsilon)\|dt. \tag{8.1}$$

We have the following theorem.

THEOREM 4. *Assume that*: (i) *Assumption* 2 *holds*, (ii) $\theta = h$ *and* (iii) $\hat{r}(\varepsilon) = o(1)$ *as* $\varepsilon \to 0$ *in* $S_{\alpha c}$. *Then*, (E_1) *is a* G.A.D.S., *where* $\|P(x,\varepsilon)\| = O(\hat{r}(\varepsilon))$ *uniformly in* $\bar{G}$ *as* $\varepsilon \to 0$ *in* $S_{\alpha c_1}$, $(0 < c \leqq c_1)$.

THEOREM 5. *Assume that*: (i) *Assumption* 2 *holds*, (ii) $\theta > h/2$, (iii) *there exists a fixed number* μ *such that* (2.5) *holds for all* $x \in G$ *and* $\varepsilon \in S_{\alpha c}$. *Then*, (E_1) *is a* G.A.D.S. *and* $\|P(x,\varepsilon)\| = O(\varepsilon^{\sigma})$ *uniformly on* $\bar{G}$ *as* $\varepsilon \to 0$ *in* $S_{\alpha c_1}$, $(0 < c_1 \leqq c)$ *where* $\sigma = \min\{\theta, 2\theta - h\}$.

These theorems improve a special case of the result of Sibuya [13].

The proofs for Theorem 4 and Theorem 5 for a fixed path Γ_x are similar to those for Theorem 1 and Theorem 2, respectively. Then, employ a method similar to that in Gingold and Hsieh [5], there exist $P(x,\varepsilon)$ globally in G satisfying these theorems.

Remark. (1) Similar to discussions in Gingold and Hsieh [5], Assumption 2 is satisfied if the following condition is satisfied:

(K) There exist two points a and b on ∂G and positive constants α_1, α_2 and $\delta(\delta < \pi/2)$ such that every x in G can be connected with a and b by a smooth Jordan curve in G

$$\Gamma_x : s = s(\tau), \qquad 0 \leqq \tau \leqq 1 \tag{8.2}$$

satisfying:

$$s(0) = a, \quad s(1) = b, \quad s(\xi) = x \qquad (0 < \xi < 1) \tag{8.3}$$

$$0 < \alpha_1 \leqq \left| \frac{ds}{d\tau} \right| \leqq \alpha_2 \quad \text{for every } \Gamma_x, \tag{8.4}$$

$$-\frac{\pi}{2} + \delta \leqq \arg\left\{ \left[d_j(s(\tau),\varepsilon) - d_k(s(\tau),\varepsilon) \right] \frac{ds}{d\tau} \right\} - h \arg \varepsilon \leqq \frac{\pi}{2} - \delta \qquad (\mathrm{mod}\, 2\pi) \tag{8.5}$$

or

$$\frac{\pi}{2} + \delta \leqq \arg\left\{ \left[d_j(s(\tau),\varepsilon) - d_k(s(\tau),\varepsilon) \right] \frac{ds}{d\tau} \right\} - h \arg \varepsilon \leqq \frac{3\pi}{2} - \delta \qquad (\mathrm{mod}\, 2\pi) \tag{8.6}$$

for every Γ_x, $0 \leqq \tau \leqq 1$, $\varepsilon \in S_{\alpha c}$, $j \neq k, j, k = 1, 2, \cdots, n$.

(2) In particular, if G is a simply connected domain bounded by two conjugate circular arcs connecting the origin $x = 0$ and $x = b$, $(b > 0)$ such that their tangents at $x = 0$ are the straight lines $\arg x = \pm\gamma$, $(0 < \gamma < \pi/2)$, respectively, and if $\mathrm{Re}\{d_j(x,\varepsilon) - d_k(x,\varepsilon)\} \neq 0$ for $x \in G$, $\varepsilon \in S_{\alpha c}$, $j \neq k, j, k = 1, 2, \cdots, n$, then, for $x \in G$, $\mathrm{Im}\, x \neq 0$, Γ_x can be taken to be the circular arc passing through $s = 0$, $s = x$ and $s = b$, and for $\mathrm{Im}\, x = 0$, Γ_x is taken to be $\overline{Ob}$. Condition (K) is satisfied if γ and α are sufficiently small.

REFERENCES

[1] W. A. COPPEL, *Dichotomy and reducibility*, J. Differential Equations, 3 (1967), pp. 500–521.

[2] A. DEVINATZ, *An asymptotic theorem for systems of linear differential equations*, Trans. Amer. Math. Soc., 160 (1971), pp. 353–363.

[3] H. GINGOLD, *A method of global block diagonalization for matrix-valued functions*, this Journal, 9 (1978), pp. 1076–1982.

[4] _____, *On the location of zeroes of oscillatory solutions of* $y^{(n)} = c(x)y$, preprint, West Virginia Univ. Morgantown.

[5] H. GINGOLD AND P. F. HSIEH, *On global simplification of a singularly perturbed system of linear ordinary differential equations*, Funk. Ekv., 25 (1982), pp. 269–281.

[6] W. A. HARRIS, JR. AND D. A. LUTZ, *A unified theory of asymptotic integration*, J. Math. Anal. Appl., 57 (1977), pp. 571–586.

[7] P. HARTMAN AND A. WINTNER, *Asymptotic integration of linear differential equations*, Amer. J. Math., 77 (1955), pp. 45–86.

[8] P. F. HSIEH, *On an analytic simplification of a system of linear differential equations containing a parameter*, Proc. Amer. Math. Soc., 19 (1968), pp. 1201–1206.

[9] T. KATO, *Notes on perturbation theory*, Bull. Amer. Math. Soc., 66 (1955), pp. 146 ff.

[10] R. Y. LEE, *Turning point problems of almost diagonal systems*, J. Math. Anal. Appl., 24 (1968), pp. 509–526.

[11] N. LEVINSON, *The asymptotic nature of solutions of linear differential equations*, Duke Math. J., 15 (1948), pp. 111–126.

[12] F. RIESZ AND B. SZ-NAGY, *Functional Analysis*, Frederick Ungar, New York, 1955.

[13] Y. SIBUYA, *Asymptotic solutions of a system of linear ordinary differential equations containing a parameter*, Funk. Ekv., 4 (1962), pp. 83–113.

[14] W. WASOW, *Asymptotic Expansions for Ordinary Differential Equations*, John Wiley, New York, 1965.

[15] _____, *On turning point problems for systems with almost diagonal coefficient matrix*, Funk. Ekv., 8 (1966), pp. 143–171.

SIAM J. MATH. ANAL.
Vol. 17. No. 1, January 1986

OSCILLATION RESULTS FOR SECOND ORDER DIFFERENTIAL SYSTEMS*

G. J. BUTLER† AND L. H. ERBE†

Abstract. Oscillation criteria are developed for the second order vector differential system (1) $y''+Q(t)y=0$ where $Q(t)$ is an $n\times n$ real continuous symmetric matrix. We show that $\lambda_1(t)=\lambda_1(\int_a^t Q(s)\,ds)\to+\infty$ as $t\to\infty$ along with either a condition on the trace of $\int_a^t Q(s)\,ds$, or a condition on the growth of $\lambda_1(t)/\lambda_n(t)$, imply oscillation of all solutions of (1). (Here $\lambda_1(\cdot)\geqq\cdots\geqq\lambda_n(\cdot)$ denote the (ordered) eigenvalues of the $n\times n$ matrix.). The results obtained generalize a theorem of Mingarelli.

1. Introduction.

Consider the second order vector differential equation

$$y''+Q(t)y=0,\qquad t\in[a,\infty) \tag{1.1}$$

where $Q(t)$ is a continuous real symmetric $n\times n$ matrix function. In partial answer to a conjecture in [6], it has been recently shown by Mingarelli [7] that the condition

$$\lim_{t\to\infty}\lambda_1\left(\int_a^t Q(s)\,ds\right)=+\infty, \tag{1.2}$$

where $\lambda_1(\cdot)$ denotes the maximum eigenvalue of the matrix, implies oscillation of (1.1) under the assumption that condition

$$\liminf_{t\to\infty}\frac{1}{t}\operatorname{tr}\left(\int_a^t Q(s)\,ds\right)>-\infty \tag{1.3}$$

holds, where $\operatorname{tr}(\cdot)$ represents the trace of the matrix. In this paper, we show that condition (1.2) implies oscillation of (1.1), even if (1.3) does not hold, provided a weaker condition than (1.3) holds (cf. §2). In addition, we show in §3 that condition (1.2) can also be weakened somewhat provided a certain relation holds between the largest and smallest eigenvalues of $\int_a^t Q(s)\,ds$ as $t\to\infty$.

Before proceeding with the statement of the results, we recall some pertinent definitions and notation which will be subsequently used. For any $n\times n$ matrix A, the *transpose* will be denoted by A^*; similarly y^* denotes the transpose of the column vector y. If t_0, $t_1\in[a,\infty)$, $t_0\neq t_1$ and if there exists a nontrivial solution $y(t)$ of (1.1) which vanishes at t_0 and t_1, then t_0 and t_1 are said to be (*mutually*) *conjugate* relative to (1.1). Equation (1.1) is said to be *disconjugate* on an interval $J\subset[a,\infty)$ if every nontrivial solution of (1.1) vanishes at most once in J and (1.1) is said to be *oscillatory* if for each $t_0>a$ there exists $t_1>t_0$ such that (1.1) is not disconjugate on $[t_0,t_1]$.

The matrix differential system associated to (1.1) is

$$Y''+Q(t)Y=0,\qquad t\in[a,\infty) \tag{1.4}$$

where Y is an $n\times n$ matrix and Q is as in (1.1). A solution of (1.4) is said to be *nontrivial* if $\det Y(t)\neq 0$ for at least one $t\in[a,\infty)$ and a nontrivial solution Y is said to be

*Received by the editors May 5, 1983, and in revised form April 5, 1984. This research was supported by the Natural Sciences and Engineering Research Council of Canada.

†Department of Mathematics, University of Alberta, Edmonton, Alberta, Canada.

prepared or *self-conjugate* in case

$$Y^*(t)Y'(t)-Y^{*\prime}(t)Y(t)\equiv 0, \qquad t\in[a,\infty). \tag{1.5}$$

(We note that for *any* solution Y of (1.4) the expression on the left side of (1.5) is constant.)

Equation (1.4) is said to be *oscillatory* in case the determinant of every nontrivial prepared solution vanishes on $[b,+\infty)$ for each $b>a$. This is equivalent to oscillation of equation (1.1) since any solution of (1.1) is of the form $y(t)\equiv Y(t)\alpha$ for some constant vector α and some nontrivial prepared solution $Y(t)$ of (1.4).

If A is a real symmetric $n\times n$ matrix, then its eigenvalues $\lambda_k(A)$, $1\leqq k\leqq n$ (which are all real) will be assumed to be ordered so that

$$\lambda_1(A)\geqq\lambda_2(A)\geqq\cdots\geqq\lambda_n(A), \tag{1.6}$$

$$\operatorname{tr}A=\sum_{k=1}^{n}\lambda_k(A). \tag{1.7}$$

We denote by $\mathcal{S}$ the linear space of all $n\times n$ real symmetric matrices. A linear functional $\phi\colon \mathcal{S}\to(-\infty,+\infty)$ is said to be *positive* if $\phi(A)\geqq 0$ for $A\in\mathcal{S}$ and $A\geqq 0$ (i.e. A symmetric positive semidefinite). Many of the recent results concerned with oscillation criteria for (1.1) and (1.4) have been based on the use of positive linear functionals. We refer to [6], [3], [4], [11], [5], [1], [10] and the references therein. The basic result obtained is that equation (1.1) (or (1.4)) is oscillatory on $[a,\infty)$ in case there exists a positive linear functional ϕ such that the scalar equation

$$u''+\phi(Q(t))u=0 \tag{1.8}$$

is oscillatory, where we assume that $\phi(I)=1$ (I=identity matrix) (cf. [3]). Hartman in [5] showed that many of the oscillation criteria can be generalized by replacing linear functionals by suitable *nonlinear* functionals.

We present in §4 a class of examples which serve to illustrate the results obtained in §§2 and 3. In particular, it is shown that there exist continuous matrix functions $Q(t)$ which are such that for *any* positive linear functional, equation (1.8) is nonoscillatory but by applying the results obtained here we may conclude that (1.1) (or (1.4)) is oscillatory. We also demonstrate that the results of §§2 and 3 are independent.

For completeness, we recall the following definitions (cf. [8]). For any subset E of the real line R, $\nu(E)$ denotes the Lebesgue measure of E. If $F(t)$ denotes a continuous real-valued function and if l, L satisfy $-\infty<l$, $L\leqq+\infty$, then we say $\operatorname{lim\,approx\,inf}_{t\to\infty}F(t)=l$ in case $\nu\{t\colon F(t)\leqq l_1\}<+\infty$ for all $l_1<l$ and $\nu\{t\colon F(t)\leqq l_2\}=+\infty$ for all $l_2>l$. Similarly, $\operatorname{lim\,approx\,sup}_{t\to\infty}F(t)=L$ in case $\nu\{t\colon F(t)\geqq L_1\}=+\infty$ for all $L_1<L$ and $\nu\{t\colon F(t)\geqq L_2\}<+\infty$ for all $L_2>L$. Finally, $\operatorname{lim\,approx}_{t\to\infty}F(t)=\lambda$ in case $\operatorname{lim\,approx\,sup}_{t\to\infty}F(t)=\operatorname{lim}_{t\to\infty}\operatorname{approx\,inf}F(t)=\lambda$. Clearly, $\operatorname{lim\,inf}_{t\to\infty}F(t)\leqq\operatorname{lim\,approx\,inf}_{t\to\infty}F(t)\leqq\operatorname{lim}_{t\to\infty}\operatorname{approx\,sup}F(t)\leqq\operatorname{lim}_{t\to\infty}\sup F(t)$.

2. We begin with the statement of the main result in this section.

THEOREM 2.1. *Let $g=g(t)$ be positive, absolutely continuous, and nondecreasing on $[a,+\infty)$ and assume*

$$\text{(i)}\quad \operatorname*{lim\,approx\,inf}_{t\to\infty}\frac{1}{g(t)}\operatorname{tr}\left(\int_a^t Q(s)\,ds\right)=l>-\infty,$$

$$\text{(ii)}\quad \operatorname*{lim\,approx\,inf}_{t\to\infty}\frac{1}{g(t)}\int_a^t\left(\lambda_1\left(\int_a^s Q(\sigma)\,d\sigma\right)\right)^2 ds=+\infty$$

and that

$$\lim_{t\to\infty} \lambda_1\Big(\int_a^t Q(s)\,ds\Big) = +\infty. \tag{1.2}$$

Then (1.1) *is oscillatory.*

The proof of Theorem 2.1 will be based on the following lemma which generalizes [7, Lemma 2.1].

LEMMA 2.2. *Let* $V(t)=V^*(t)$ *be a continuous* $n\times n$ *symmetric matrix function and suppose that*

$$\operatorname*{lim\,approx\,sup}_{t\to\infty} \frac{\operatorname{tr}\big(V(t)+\int_a^t V^2(s)\,ds\big)}{g(t)} = L < +\infty \tag{2.1}$$

where g *is positive, absolutely continuous, and nondecreasing on* $(a, +\infty)$. *Then it follows that*

$$\operatorname*{lim\,approx\,inf}_{t\to\infty} \frac{1}{g(t)} \int_a^t \lambda_1\big(V^2(s)\big)\,ds < +\infty. \tag{2.2}$$

Proof. Since $V(t)=V^*(t)$, it follows that $V^2(t)\geqq 0$ and so $F(t)\equiv\int_a^t V^2(s)\,ds$ satisfies $F(t_2)\geqq F(t_1)$, $a\leqq t_1<t_2<+\infty$ (i.e., $F(t_2)-F(t_1)$ is nonnegative definite). Therefore, since $(\lambda_1(V(t)))^2\leqq\lambda_1(V^2(t))\leqq\operatorname{tr}(V^2(t))\leqq n\lambda_1(V^2(t))$, it follows that if

$$\operatorname*{lim\,approx\,inf}_{t\to\infty} \frac{1}{g(t)} \int_a^t \lambda_1\big(V^2(s)\big)\,ds = +\infty$$

then

$$\operatorname*{lim\,approx\,inf}_{t\to\infty} \frac{1}{g(t)} \int_a^t \operatorname{tr}\big(V^2(s)\big)\,ds = \operatorname*{lim\,approx\,inf}_{t\to\infty} \frac{1}{g(t)} \operatorname{tr}\Big(\int_a^t V^2(s)\,ds\Big) = +\infty.$$

We define

$$H(t) \equiv \operatorname{tr}\Big(\int_a^t V^2(s)\,ds\Big) - Mg(t), \qquad t\geqq a$$

where $M>L$ is arbitrary. It follows then that

$$\operatorname*{lim\,approx\,inf}_{t\to\infty} \frac{H(t)}{g(t)} = +\infty$$

so that the inequality $H(t)>g(t)$ holds on a set E with $\nu(E)=+\infty$. Also, if we define the set E_1 by

$$E_1 \equiv \Big\{t:\ \frac{1}{g(t)}\big(\operatorname{tr} V(t)+H(t)\big)\geqq 0\Big\},$$

then by (2.1) we see that $\nu(E_1)<\infty$ since $M>L$ and so the complement of E_1 relative to $[a, +\infty)$, which we denote by $\tilde{E}_1$, satisfies $\nu(\tilde{E}_1)=+\infty$, and $H(t)<-\operatorname{tr} V(t)$ on $\tilde{E}_1$. Since $\nu(E\cap\tilde{E}_1)=+\infty$, we see that the inequality $g(t)<H(t)<-\operatorname{tr} V(t)$ holds on $E\cap\tilde{E}_1$. Thus, $g^2(t)\leqq H^2(t)\leqq(\operatorname{tr} V(t))^2\leqq n\operatorname{tr}(V^2(T))=n(H'(t)+Mg'(t))$ holds (a.e.) on $E\cap\tilde{E}_1$ and so we have

$$\frac{1}{n}\leqq\frac{H'(t)}{H^2(t)}+\frac{Mg'(t)}{H^2(t)}\leqq\frac{H'(t)}{H^2(t)}+\frac{Mg'(t)}{g^2(t)} \quad \text{on } E\cap\tilde{E}_1. \tag{2.3}$$

This is a contradiction since the integral of the right-hand side of (2.3) is less than $\int_{H(a)}^{\infty} u^{-2}\,du + \int_{g(a)}^{\infty} Mu^{-2}\,du$ which is finite, whereas the integral of the left-hand side is infinite since $\nu(E \cap \tilde{E_1}) = +\infty$. This proves the lemma.

Proof of Theorem 2.1. If (1.4) is not oscillatory, then there exists a nontrivial prepared solution $Y(t)$ of (1.4) with $\det Y(t) \neq 0$ on $[t_0, \infty)$ for some $t_0 \geqq a$. Let $V(t) = Y'(t)Y^{-1}(t)$, $t \geqq t_0$ so that $V^*(t) = V(t)$ and $V(t)$ satisfies the matrix Riccati equation

$$-V'(t) = Q(t) + V^2(t), \qquad t \geqq t_0. \tag{2.4}$$

Integrating (2.4) we obtain the equivalent integral equation

$$-V(t) + V(t_0) = \int_{t_0}^{t} Q(s)\,ds + \int_{t_0}^{t} V^2(s)\,ds \tag{2.5}$$

and hence

$$-V(t) + V(t_0) \geqq \int_{t_0}^{t} Q(s)\,ds. \tag{2.6}$$

Therefore, $\lambda_1(\int_{t_0}^{t} Q(s)\,ds) \leqq \lambda_1(-V(t)) + \lambda_1(V(t_0))$ by the subadditivity of λ_1 and (2.6). Since $\lambda_1(V^2(t)) \geqq (\lambda_1(-V(t))^2)$ we see that $\lim_{t\to\infty} \lambda_1(-V(t)) = \lim_{t\to\infty} \lambda_1(V^2(t)) = +\infty$. Now by condition (i) of the Theorem 2.1 and (2.5) it follows that

$$\begin{aligned} &\limapprox\sup_{t\to\infty} \left(\frac{1}{g(t)} \left(\operatorname{tr}\left(V(t) + \int_{t_0}^{t} V^2(s)\,ds \right) \right) \right) \\ &= \limapprox\sup_{t\to\infty} \left(\frac{1}{g(t)} \left(\operatorname{tr}\left(V(t_0) - \int_{t_0}^{t} Q(s)\,ds \right) \right) \right) = k - l < +\infty \end{aligned}$$

where $k = \lim_{t\to\infty} (\operatorname{tr} V(t_0)/g(t))$. Therefore, by Lemma 2.2 we have that

$$\limapprox\inf_{t\to\infty} \frac{1}{g(t)} \int_{t_0}^{t} \lambda_1(V^2(s))\,ds < +\infty.$$

Again, from (2.6) we have $\lambda_1(\int_{t_0}^{t} Q(s)\,ds) \leqq 2\lambda_1(-V(t))$ for all large t (since $\lambda_1(-V(t)) \to +\infty$) and therefore

$$\left(\lambda_1\left(\int_{t_0}^{t} Q(s)\,ds \right) \right)^2 \leqq 4\left(\lambda_1(-V(t))^2 \right) \leqq 4\lambda_1(V^2(t)) \tag{2.7}$$

for all large t. Thus, we obtain by integration

$$\frac{1}{g(t)} \int_{t_0}^{t} \left(\lambda_1\left(\int_{t_0}^{s} Q(r)\,dr \right) \right)^2 ds \leqq \frac{4}{g(t)} \int_{t_0}^{t} \lambda_1(V^2(s))\,ds \tag{2.8}$$

for all large r and hence taking lim approx inf of both sides as $t \to \infty$ we have a contradiction to condition (ii) of the hypotheses. This proves Theorem 2.1.

Remark. If $g(t) = t$ in Theorem 2.1, then condition (1.2) implies that

$$\lim_{t\to\infty} \frac{1}{t} \int_{a}^{t} \left(\lambda_1\left(\int_{a}^{s} Q(r)\,dr \right) \right)^2 ds = +\infty$$

so that Theorem 2.1 includes the results of Mingarelli [7] with "lim inf" replaced by "lim approx inf". In §4, we investigate further conditions (i) and (ii) by means of several examples.

Theorem 2.1 may also be generalized by considering the principal submatrices of $\int_a^t Q(s)\,ds$. We recall the notation (cf. [2]): For any $n\times n$ symmetric matrix A, the sequence of symmetric matrices $A_k\equiv(a_{ij})$, $i,j=1,\cdots,k$, for $k=1,2,\cdots,n$ satisfies $\lambda_{j+1}(A_{k+1})\leqq\lambda_j(A_k)\leqq\lambda_j(A_{k+1})$ where $\lambda_j(A_k)$ denotes the jth characteristic root of A_k (cf. [2, p. 113]). We may then state the next corollary.

COROLLARY 2.3. *Let $g=g(t)$ be positive, continuous, and nondecreasing in $[a,+\infty)$ and assume there exists k, $1\leqq k\leqq n$ such that*

$$(\mathrm{i})_k \quad \liminf_{t\to\infty}\operatorname{approx}\frac{1}{g(t)}\operatorname{tr}\left(\int_a^t Q_k(s)\,ds\right)>-\infty,$$

$$(\mathrm{ii})_k \quad \liminf_{t\to\infty}\operatorname{approx}\frac{1}{g(t)}\int_a^t\left(\lambda_1\left(\int_a^s Q_k(r)\,dr\right)\right)^2 ds=+\infty$$

and that

$$(1.2)_k \qquad \lim_{t\to\infty}\lambda_1\left(\int_a^t Q_k(s)\,ds\right)=+\infty.$$

Then equation (1.1) *is oscillatory.*

Proof. The proof proceeds as in Theorem 2.1 to obtain equation (2.5) from which we have

$$(2.9) \qquad -V_k(t)+V_k(t_0)=\int_{t_0}^t Q_k(s)\,ds+\int_{t_0}^t\left(V^2(s)\right)_k ds\geqq\int_{t_0}^t Q_k(s)\,ds,$$

and therefore $\lambda_1(-V_k(t))\to+\infty$ as $t\to\infty$ by $(1.2)_k$. A straightforward modification of the proof of Theorem 2.1 now yields a contradiction to condition (ii) and this proves Corollary 2.3.

3. In this section, we relax the condition (1.2) and replace hypotheses (i) and (ii) of Theorem 2.1 by a condition on the relative rates of growth of the largest and smallest eigenvalues of $\int_a^t Q(s)\,ds$ as $t\to\infty$. Abbreviate the notation for the eigenvalues of $\int_a^t Q(s)\,ds$ to $\lambda_1(t)\geqq\lambda_2(t)\geqq\cdots\geqq\lambda_n(t)$.

We shall require the following simple lemma.

LEMMA 3.1. *Let $p(t)$ be locally bounded, nonnegative and measurable on $[a,\infty)$ with $p(t)$ not zero a.e. Let $q(t)$ be nonnegative and locally integrable, such that*

$$p(t)\geqq q(t)\int_a^t p^2(s)\,ds \quad \textit{for almost all } t\geqq a.$$

Then for all sufficiently large $\bar a\geqq a$, we have $q\in L^2[\bar a,\infty)$.

Proof of Lemma 3.1. Let $P(t)=\int_a^t p^2(s)\,ds$. Then $P(t)$ is absolutely continuous, and by hypothesis, $P(t)>0$ for $t>a^*$, say. We have $P'(t)=P^2(t)\geqq q^2(t)P^2(t)$, and integrating $P'(s)/P^2(s)$ from $\bar a>a^*$ to t, we obtain

$$\frac{1}{P(\bar a)}-\frac{1}{P(t)}=\int_{\bar a}^t q^2(s)\,ds,$$

and so

$$\int_{\bar a}^{\infty} q^2(s)\,ds\leqq\frac{1}{P(\bar a)},$$

proving the lemma.

THEOREM 3.2. *Let one of the following set of hypotheses hold*:

(1) (a) $\operatorname{lim\,approx\,inf}\lambda_1(t)=\infty$ *and*

(b) $\operatorname{lim\,approx\,sup}_{t\to\infty}|\lambda_1(t)/\lambda_n(t)|>0$;

or

(2) (a) $\operatorname{lim\,approx\,sup}_{t\to\infty}\lambda_1(t)=\infty$ *and*

(b) $\operatorname{lim\,approx\,inf}_{t\to\infty}|\lambda_1(t)/\lambda_n(t)|>0$.

Then (1.1) *is oscillatory.*

Proof. Assume that (1) or (2) holds and suppose that (1.1) has a nonoscillatory solution $Y(t)$. Let $W(t)=-Y'(t)Y^{-1}(t)$ so that $W(t)$ satisfies

$$W(t)-\int_{t_0}^{t} W^2(s)\,ds=\int_a^t Q(s)\,ds+C, \qquad t\geqq t_0, \tag{3.1}$$

where $C=W(t_0)-\int_a^{t_0}Q(s)\,ds$.

It is known that for any continuous are symmetric matrix-valued function, a continuously varying orthonormal system may be selected [9]. It follows that we may choose a locally integrable vector $x(t)$ with $\|x(t)\|=1$ such that

$$x^*(t)\left(\int_a^t Q(s)\,ds\right)x(t)=\lambda_1(t). \tag{3.2}$$

Let the eigenvalues of $W(t)$ be $\mu_1(t)\geqq\cdots\geqq\mu_n(t)$. By the preceding remark, we may select a system of (orthonormal) locally integrable eigenvectors $e_i(t)$, such that

$$W(t)e_i(t)=\mu_i(t)e_i(t), \qquad e_i^*(t)e_j(t)=\delta_{ij}. \tag{3.3}$$

Define the functions $c_i(s,t)$, $i=1,\cdots,n$, $a\leqq s$, $t<\infty$, by

$$x(t)=\sum_{i=1}^{n} c_i(s,t)e_i(s). \tag{3.4}$$

The $c_i(s,t)$ are projections of $x(t)$ on to the orthonormal system $\{e_i(s)\}_{i=1}^n$, and are locally integrable with respect to both s and t.

We have

$$W(s)x(t)=\sum_{i=1}^{n}\mu_i(s)c_i(s,t)e_i(s). \tag{3.5}$$

Denote the left-hand side of (3.1) by $\Phi(t)$. From (3.1) to (3.5), we have

$$x^*(t)\Phi(t)x(t)=\sum_{i=1}^{n}\mu_i(t)c_i^2(t,t)-\int_{t_0}^{t}\left(\sum_{i=1}^{n}\mu_i^2(s)c_i^2(s,t)\right)ds. \tag{3.6}$$

At this point, in order to give a clearer presentation of the argument, we shall concentrate on the case $n=2$. Introduce angle-functions ϕ,θ by defining

$$\begin{aligned} e_1(s)&=\cos\phi(s)e_1(a)+\sin\phi(s)e_2(a),\\ e_2(s)&=-\sin\phi(s)e_1(a)+\cos\phi(s)e_2(a); \end{aligned} \tag{3.7}$$

$$c_1(a,t)=\cos\theta(t), \qquad c_2(a,t)=\sin\theta(t). \tag{3.8}$$

From (3.7) and (3.8), we find that

$$c_1(s,t)=\cos(\theta(t)-\phi(s)),\ c_2(s,t)=\sin(\theta(t)-\phi(s)).$$

If we put $\alpha(t)=\theta(t)-\phi(t)$, we may write (3.6) as

$$x^*(t)\Phi(t)x(t)=\mu_1(t)\cos^2\alpha(t)+\mu_2(t)\sin^2\alpha(t)$$

(3.9)

$$-\int_{t_0}^{t}\mu_1^2(s)\cos^2\{\alpha(t)+\phi(t)-\phi(s)\}\,ds-\int_{t_0}^{t}\mu_2^2(s)\sin^2\{\alpha(t)+\phi(t)-\phi(s)\}\,ds.$$

Now $x^*(t)\Phi(t)x(t)=\lambda_1(t)+x^*(t)Cx(t)$. Since C is constant and $\|x(t)\|=1$, we have

(3.10.1) $$\operatorname*{lim\,approx\,inf}_{t\to\infty} x^*(t)\Phi(t)x(t)=\infty$$

or

(3.10.2) $$\operatorname*{lim\,approx\,sup}_{t\to\infty} x^*(t)\Phi(t)x(t)=\infty$$

according as hypothesis (1a) or (2a) holds.

Our object is to demonstrate the incompatibility of (3.9) and (3.10.1) or (3.10.2). Parts b) of hypotheses (1) and (2) imply that there exists δ with $0<\delta<1$ such that

(3.11) $$\left|\frac{\lambda_1(t)}{\lambda_2(t)}\right|>\delta \quad \text{for all } t\in T_1$$

where $T_1\subset[a,\infty)$ satisfies

(3.12.1) $$\nu(T_1)=\infty \quad \text{(if (1) holds)},$$

(3.12.2) $$\nu(\tilde{T}_1)<\infty \quad \text{(if (2) holds)},$$

with $\tilde{T}_1=[a,\infty)-T_1$.

Let $\lambda(t)=\max(|\lambda_1(t)|,|\lambda_2(t)|)$. For $t\in T_1$, we have $\lambda(t)\leqq\lambda_1(t)/\delta$. For any $y\in R^n$ with $\|y\|\leqq\frac{1}{4}\delta$, we have

(3.13)
$$\begin{aligned}(x(t)+y)^*\Phi(t)(x(t)+y)&=x(*t)\Phi(t)x(t)+y^*\Phi(t)y+x^*(t)\Phi(t)y+y^*\Phi(t)x(t)\\&\geqq\lambda_1(t)-2\lambda(t)\|y\|-\lambda(t)\|y\|^2-\left(1+\frac{1}{4}\delta\right)^2\|C\|\\&\geqq\lambda_1(t)-\frac{1}{2}\lambda_1(t)-\frac{1}{16}\delta\lambda_1(t)-\left(1+\frac{1}{4}\delta\right)^2\|C\|,\quad \text{if } t\in T_1\\&\geqq\frac{1}{3}\lambda_1(t)-2\|C\|.\end{aligned}$$

Recall that $x(t)=\cos\theta(t)e_1(a)+\sin\theta(t)e_2(a)$. Simple trigonometry shows that we may choose $y=\hat{y}(t)$ with $\|\hat{y}(t)\|\leqq\frac{1}{4}\delta$, such that

$$x(t)+\hat{y}(t)=\cos\theta(t)e_1(a)+\sin\hat{\theta}(t)e_2(a)$$

where $\hat{\theta}(t)=\theta(t)+\delta/4$.

Now define $y(t)$ by

$$y(t)=\begin{cases}0 & \text{if } |\alpha(t)|\geqq\frac{\delta}{8}(\text{mod}\,\pi) \text{ and } \left|\alpha(t)-\frac{\pi}{2}\right|\geqq\frac{\delta}{8}(\text{mod}\,\pi),\\ y(t) & \text{otherwise,}\end{cases}$$

and let $\hat{x}(t)=x(t)+y(t)$. If $\hat{\alpha}(t)$ is defined by

$$\hat{x}(t)=\cos\hat{\theta}(t)e_1(a)+\sin\hat{\theta}(t)e_2(a), \qquad \hat{\alpha}(t)=\hat{\theta}(t)-\phi(t),$$

we obtain

$$|\hat{\alpha}(t)|\geqq\frac{\delta}{8}(\text{mod}\,\pi), \qquad \left|\hat{\alpha}(t)-\frac{\pi}{2}\right|\geqq\frac{\delta}{8}(\text{mod}\,\pi) \tag{3.14}$$

for all $t\geqq a$. Now (3.9) and (3.13) give

$$\begin{aligned}\hat{x}^*(t)\Phi(t)\hat{x}(t)&=\mu_1\cos^2\hat{\alpha}(t)+\mu_2(t)\sin^2\hat{\alpha}(t)\\ &\quad-\int_a^t\mu_2^2(s)\cos^2\{\hat{\alpha}(t)+\phi(t)-\phi(s)\}\,ds\\ &\quad-\int_a^t\mu_2^2(s)\sin^2\{\alpha(t)+\phi(t)-\phi(s)\}\,ds\\ &\geqq\frac{1}{3}\lambda_1(t)-2\|C\|, \qquad t\in T_1.\end{aligned} \tag{3.15}$$

Now parts (a) of the hypotheses of the theorem yield the existence of a subset T_2 of $[a,\infty)$ such that

$$\frac{1}{3}\lambda_1(t)-2\|C\|>0, \qquad t\in T_2 \tag{3.16}$$

and

$$\nu(\tilde{T}_2)<\infty \quad \text{(if (1) holds)}, \tag{3.17.1}$$

$$\nu(T_2)=\infty \quad \text{(if (2) holds)}, \tag{3.17.2}$$

($\tilde{T}_2=[a,\infty)-T_2$). Now (3.12), (3.14), (3.15), (3.16) and (3.17) show that there exists a subset T ($=T_1\cap T_2$) of $[a,\infty)$ with $\nu(T)=\infty$, such that

$$\begin{aligned}&\mu_1(t)\cos^2\hat{\alpha}(t)+\mu_2(t)\sin^2\hat{\alpha}(t)-\int_a^t\mu_1^2(s)\cos^2\{\hat{\alpha}(t)+\phi(t)-\phi(s)\}\,ds\\ &\qquad-\int_a^t\mu_2^2(s)\sin^2\{\hat{\alpha}(t)+\phi(t)-\phi(s)\}\,ds>0, \qquad t\in T,\end{aligned} \tag{3.18}$$

where $\alpha(t)$ is bounded away from 0 and $\pi/2$ (mod π). Let

$$U_1(t)=\mu_1(t)\cos^2\hat{\alpha}(t)-\int_a^t\mu_1^2(s)\cos^2\{\hat{\alpha}(t)+\phi(t)-\phi(s)\}\,ds \tag{3.19}$$

and let $S_1=\{t\colon U_1(t)>0\}$. By (3.14), there exist $\delta_1>0$ and a positive integer m such that

$$\cos^2(\hat{\alpha}(t)+\omega)\geqq\delta_1 \quad \text{for } t\geqq a, \qquad |\omega|\leqq\frac{1}{m} \quad (\text{mod}\,\pi). \tag{3.20}$$

For $i=1,2,\cdots,m$, define

$$A_i=\left\{t\geqq a:\ \frac{i-1}{m}\leqq\phi(t)<\frac{i}{m}(\mathrm{mod}\,\pi)\right\},$$

and let

$$\hat{\mu}_i(t)=\begin{cases}|\mu_1(t)| & t\in A_i,\\ 0, & \text{otherwise.}\end{cases}$$

If $t\in A_i$, we have by (3.20) that

$$U_1(t)\leqq\hat{\mu}_i(t)-\delta_1\int_a^t\hat{\mu}_i^2(s)\,ds. \tag{3.21}$$

By Lemma 3.1, either $\mu_i(t)=0$ a.e. or, if $q_i(t)$ is defined by

$$\hat{\mu}_i(t)=\hat{q}_i(t)\int_a^t\hat{\mu}_i^2(s)\,ds,$$

we have $\hat{q}_i\in L^2[\bar{a},\infty)$ for $\bar{a}$ sufficiently large. In the latter case, the set of t for which $\hat{q}_i(t)\geqq\delta_1$ has finite measure, and in either case, the set of t in A_i for which $U_1(t)>0$, has finite measure. Since $[a,\infty)=\bigcup_{i=1}^m A_i$, we have $\nu(S_1)<\infty$. Similar reasoning shows that if

$$U_2(t)=\mu_2(t)\sin^2\hat{\alpha}(t)-\int_a^t\mu_2^2(s)\sin^2\{\hat{\alpha}(t)+\phi(t)-\phi(s)\}\,ds$$

and $S_2=\{t:\ U_2(t)>0\}$, then $\nu(S_2)<\infty$. But this is in contradiction to (3.18), and now the theorem is proved for the case $n=2$.

In the general case $n>2$, the basic idea of proof is the same. We introduce the orthogonal matrix $U(s)$ whose rows are the eigenvectors $e_i(s)$ of $W(s)$ (see (3.3)). Denote the vector $(c_1(s,t),c_2(s,t),\cdots,c_n(s,t))^*$ by $c(s,t)$. Assuming, without loss of generality, that $U(a)$ is the identity matrix, (3.4) gives

$$c(s,t)=U^{-1}(s)c(a,t). \tag{3.22}$$

Now denote $U^{-1}(s)$ by $V(s)$, the components of $V(t)c(a,t)$ by $v_i(t)$ and the components of $(V(s)-V(t))c(a,t)$ by $w_i(s,t)$, $i=1,\cdots,n$. From (3.6), we may write $x^*(t)\Phi(t)x(t)$ as

$$\sum_{i=1}^n\mu_i(t)v_i^2(t)=\int_a^t\sum_{i=1}^n\mu_i^2(s)\big(v_i(t)+w_i(s,t)\big)^2ds. \tag{3.23}$$

By replacing $x(t)$ by an appropriate $\hat{x}(t)$ and obtaining the corresponding functions $\hat{v}_i(t)$, $\hat{w}_i(s,t)$, we have

$$\hat{x}^*(t)\Phi(t)\hat{x}(t)=\sum_{i=1}^n\mu_i(t)\hat{v}_i^2(t)-\int_a^t\sum_{i=1}^n\mu_i^2(s)\{\hat{v}_i(t)+\hat{w}_i(s,t)\}^2ds>0, \tag{3.24}$$

$$t\in T,$$

where $|v_i(t)|\geqq\delta_0>0$ on T, and T is a subset of $[a,\infty)$, of infinite measure, $\hat{w}_i(s,t)$ is the ith component of $(V(s)-V(t))\hat{c}(a,t)$, and $\|\hat{c}(a,t)\|=1$. Note that (3.24) is the analogue of (3.15) and (3.18). Since the $n\times n$ orthogonal matrices (identified with S^{n-1}) form a compact set, we may find a finite decomposition of S^{n-1}, $\{I_k\}_{k=1}^m$ say, such that $G,H\in I_k\Rightarrow\|G-H\|<\delta_0/2$, $k=1,\cdots,m$. Now define $A_k=\{t\geqq a:\ V(t)\in I_k\}$.

Then we shall have

$$\{\hat{v}_j(t)+\hat{w}_j(s,t)\}^2 \geqq \frac{\delta_0^2}{4} \tag{3.25}$$

whenever $s, t \in A_k$, $k=1,\cdots,m$, $j=1,\cdots,n$.

With (3.24), (3.25) and Lemma 3.1, we may complete the proof as in the case $n=2$.

4. We illustrate the results of §§2 and 3 with some examples.

1. Let $\delta > \eta > 0$ and let $\sigma=(\eta+\delta)/2-1$, $k=(\eta+\delta)/2$. Define $Q(t)$ to be

$$\begin{pmatrix} 0 & -kt^\sigma \\ -kt^\sigma & -q(t) \end{pmatrix}$$

where $q(t)=\delta t^{\delta-1}-\eta t^{\eta-1}$. Then $\int_1^t(\operatorname{tr} Q)\,ds = t^\eta - t^\delta \to -\infty$ as $t\to\infty$ and $\lambda_1(\int_1^t Q\,ds) = t^\eta \to \infty$ as $t\to\infty$. Let $g(t)=[\int_1^t(\lambda_1(\int_1^s Q\,d\sigma))^2 ds]^\gamma$ where $0<\gamma<1$. Then $g(t)\uparrow\infty$ as $t\to\infty$, and $g(t)\sim t^{(2\eta+1)\gamma}$. We have

$$\frac{\int_1^t(\operatorname{tr} Q)\,ds}{g(t)} \sim \frac{t^\eta - t^\delta}{t^{(2\eta+1)\gamma}} \geqq -t^{\delta-\gamma(2\eta+1)}$$

which is bounded below as $t\to\infty$ provided that $\delta(2\eta+1)\geqq\gamma$. Now

$$\frac{1}{g(t)}\int_1^t\left(\lambda\left(\int_1^s Q\,d\sigma\right)\right)^2 ds = \left[\int_1^t\left(\lambda_1\left(\int_1^s Q\,d\sigma\right)\right)^2 ds\right]^{1-\gamma} \to \infty$$

as $t\to\infty$. By Theorem 2.1, we shall have oscillation if $0<\eta<\delta<2\eta+1$ (choose $\gamma=\delta/2\eta+1$). If $\delta>1$, we have $1/t\int_1^t(\operatorname{tr} Q)\,ds\to-\infty$, so we cannot obtain oscillation by Mingarelli's result.

We recall that any positive functional $\phi(A)$ has the form

$$\phi(A)=\sum_{i,j=1}^n \alpha_{ij}a_{ij}$$

where $A=(a_{ij})$, and $\alpha=(\alpha_{ij})$ is nonnegative definite [11]. We have $\phi(Q(t))=-2kt^\sigma\alpha_{12}+\alpha_{22}(\eta t^{\eta-1}-\delta t^{\delta-1})\leqq 0$ for t sufficiently large, since $\alpha_{22}\geqq 0$ (with equality possible only if $\alpha_{12}=0$) and $\delta-1>\sigma>\eta-1$. Thus the 2nd order scalar equation $y''+\phi(Q(t))y=0$ is nonoscillatory for all positive functionals ϕ, and this class of tests for oscillation cannot be used.

2. To illustrate Theorem 3.2, we define the 3×3 matrix $Q(t)$ by specifying its integral from 0 to t. Let $\int_0^t Q(s)\,ds=$

$$\begin{pmatrix} t^{1/2}\cos^2 t - t^n\sin^2 t(1+\cos t) & \frac{1}{2}t^{1/2}\sin 2t\{1+t^{n-1}(1+\cos t)\} & 0 \\ \frac{1}{2}t^{1/2}\sin 2t\{1+t^{n-1}(1+\cos t)\} & t^{1/2}\sin^2 t - t^n\cos^2 t(1+\cos t) & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

where $n\geqq 1$.

A straightforward computation reveals that the eigenvalues of $\int_0^t Q(s)\,ds$ are $\lambda_1(t)=t^{1/2}$, $\lambda_2(t)=0$, $\lambda_3(t)=-t^n(1+\cos t)$. Let $t_m=(2m+1)\pi$ and $t=t_m+s$. Then it is easily verified that for t near t_m, $|\lambda_3(t)|\leqq c_1 s^2 m^n$, for some constant c_1. Since $|\lambda_1(t)|\geqq$

$c_2 m^{1/2}$, for some constant c_2, for t near t_m, we have $|\lambda_1(t)/\lambda_3(t)| \geqq 1$ if $|s| \leqq c_3 m^{1/4-n/2} (c_3 = (c_2/c_1)^{1/2})$. Thus $\{t\colon |\lambda_1(t)/\lambda_3(t)| \geqq 1\}$ has measure at least $c_4 \Sigma_{m=1}^{\infty} m^{1/4-n/2}$, for some constant $c_4 > 0$, i.e. has infinite measure if $n \leqq \frac{5}{2}$.

It follows that hypothesis (1) of Theorem 3.2 holds and equation (1.1) is oscillatory. It is easily verified that Theorem 2.1 cannot be used to verify oscillation in the case that $2 \leqq n \leqq \frac{5}{2}$.

REFERENCES

[1] W. ALLEGRETTO AND L. ERBE, *Oscillation criteria for matrix differential inequalities*, Canad. Math. Bull. 16 (1973), pp. 5–10.

[2] R. BELLMAN, *Introduction to Matrix Analysis*, 2nd ed., McGraw-Hill, New York, 1970.

[3] G. J. ETGEN AND R. T. LEWIS, *Positive functionals and oscillation criteria for differential systems* in Optimal Control and Differential Equations, A. B. Schwarzkopf, W. E. Kelley, and S. B. Eliason, eds., Academic Press, New York, 1978, pp. 245–275.

[4] G. J. ETGEN AND J. F. PAWLOWSKI, *Oscillation criteria for second order self-adjoint differential systems*, Pacific J. Math. 66 (1976), pp. 99–110.

[5] P. HARTMAN, *Oscillation criteria for self-adjoint second order differential systems and Principal sectional curvatures*, J. Differential Equations, 34 (1979), pp. 326–338.

[6] D. B. HINTON AND R. T. LEWIS, *Oscillation theory for generalized second-order differential equations*, Rocky Mountain J. Math., 10 (1980), pp. 751–766.

[7] A. B. MINGARELLI, *On a conjecture for oscillation of second order ordinary differential systems*, Proc. Amer. Math. Soc., 82 (1981), pp. 593–598.

[8] C. OLECH, Z. OPIAL AND T. WAZEWSKI, *Sur le problème d'oscillation des intégrales de l'équation* $y'' + g(t)y = 0$, Bull. de l'Académie Polonaise des Sciences, Cl. III, 5 (1957), pp. 621–626.

[9] B. N. PARLETT, *The Symmetric Eigenvalue Problem*, Prentice-Hall, Englewood Cliffs, NJ, 1980.

[10] E. C. TOMASTIK, *Oscillation of systems of second order differential equations*, J. Differential Equations, 9 (1971), pp. 436–442.

[11] T. WALTERS, *A characterization of positive linear functionals and oscillation criteria for matrix differential equations*, Proc. Amer. Math. Soc., 78 (1980), pp. 198–202.

SIAM J. MATH. ANAL.
Vol. 17, No. 1, January 1986

004

COMPARISON THEOREMS FOR CERTAIN DIFFERENTIAL SYSTEMS OF ARBITRARY ORDER*

E. C. TOMASTIK†

Abstract. Comparison theorems are given for $(k,n-k)$-focal points of systems of the form $[rx^{(k)}]^{(n-k)}-(-1)^{n-k}px=0$, where r and p are $m\times m$ matrices. The system is not assumed to be selfadjoint, and for the "comparing" equation no sign assumptions are made on the elements of $r(t)$ and $p(t)$.

1. Introduction. In this paper are given comparison theorems for $(k,n-k)$-focal points of the two nth order differential systems

$$[r(t)x^{(k)}]^{(n-k)}-(-1)^{n-k}p(t)x=0 \tag{1}$$

and

$$[R(t)y^{(k)}]^{(n-k)}-(-1)^{n-k}P(t)y=0, \tag{2}$$

where $r(t)$, $p(t)$, $R(t)$ and $P(t)$ are all $m\times m$ matrices of continuous functions, r and $R\in C^{n-k}$, and $r^{-1}(t)$ and $R^{-1}(t)$ exist on the intervals under consideration. In addition, $R^{-1}(t)$ and $P(t)$ are assumed to satisfy a certain "positivity" condition with respect to a certain cone. However, no further conditions are given on the matrices $r(t)$ and $p(t)$. In particular, $r(t)$ and $p(t)$ may have oscillatory components. Since no assumptions are made concerning the symmetry of any of the matrices $r(t)$, $p(t)$, $R(t)$ and $P(t)$, and no assumptions are made on the integers k and n, the systems (1) and (2) may or may not be selfadjoint. But even in the case that both systems are selfadjoint, the results presented here are new.

Focal points play a critical role in variational theory when the systems (1) and (2) are selfadjoint, and certain comparison theorems have long been known in this case. For example, Morse [5] gave such comparison theorems for second order systems, and Reid [8, p. 356] gives such a comparison theorem for general selfadjoint systems of order $2n$. Roughly, these results in the selfadjoint case state that if the matrices $r(t)$ and $R(t)$ are positive definite and if the matrices $P(t)-p(t)$ and $R^{-1}(t)-r^{-1}(t)$ are both positive semidefinite everywhere and one of them is positive definite at one point, then the focal point of (1) lies to the right of the focal point of (2). (In the selfadjoint case the matrices $r(t)$, $R(t)$, $p(t)$ and $P(t)$ must all be symmetric.)

In the selfadjoint case, it is thus natural to define $P>p$ if $P(t)-p(t)$ is positive semidefinite everywhere and positive definite at at least one point. Thus we can say that if P is "larger" than p and if R^{-1} is "larger" then r^{-1}, then the focal point of (1) lies to the right of the focal point of (2), in agreement with the Sturm theory for scalar second order differential equations. Since in this paper none of the matrices $r(t)$, $p(t)$, $R(t)$ and $P(t)$ are assumed to be symmetric, the notion of positive definiteness cannot be used. A natural alternative definition of $P>p$ (or $p<P$) will mean that if $p=(p_{ij})$ and $P=(P_{ij})$, then $p_{ij}(t)\le P_{ij}(t)$, and for each row strict inequality occurs for one element at at least one point. This notion is of course independent of the notion of positive definite, even if p and P are symmetric. However, a selfadjoint example is given later to

* Received by the editors January 25, 1983, and in revised form February 3, 1984.
† Department of Mathematics, University of Connecticut, Storrs, Connecticut 06268.

show that if $p<P$ under this latter definition, the focal point of (1) may be the *left* of the focal point of (2), a very surprising situation. It will be shown in this paper that a sufficient condition to assure that the focal point of (1) lies to the right of that of (2) is to assume $\int_a^t |r_{ij}^{-1}(s)|\,ds \le \int_a^t |R_{ij}^{-1}(s)|\,ds$ and $|p_{ij}(t)| \le |P_{ij}(t)|$, and that the latter inequalities become strict inequalities for at least one element at at least one point in every row. The term r_{ij}^{-1} represents the element in the ith row and the jth column of the matrix r^{-1}. The integral condition on r^{-1} and R^{-1} is of course more general than just the condition $|r_{ij}^{-1}(t)| \le |R_{ij}^{-1}(t)|$, and is thus of some interest. It is interesting to note that in the classical results in the selfadjoint case, the assumption that $r(t)$ and $R(t)$ are at least positive semidefinite is essential. However, in this paper the matrices $r(t)$ and $R(t)$ need not be positive semidefinite even in the selfadjoint case.

Comparison theorems for focal points were given by the author [9] for $m=2$, $k=1$, but they required that $P_{ij}(t) \ge 0$ and $R_{ij}^{-1}(t) \ge 0$. These conditions are replaced with the much less restrictive conditions that $P(t)$ and $R^{-1}(t)$ be positive with respect to a certain cone. Thus the results presented here are new for the second order case also. Focal points were also studied by Keener and Travis [3] for second order systems with $r(t)$ and $R(t)$ the identity and without the assumption that $P(t)$ is symmetric. In the scalar case, comparison theorems of the "integral type" were first established by Nehari [6] for second order equations, by Travis [10] for selfadjoint equations of order $2n$, by Nehari [7], Gentry and Travis [2] and Keener and Travis [4] for a somewhat more general scalar equation than (1). Elias [1] also studied focal points of scalar equations.

Throughout this paper it is assumed that some partition $\{I,J\}$ of the integers $\{1,2,\cdots,m\}$ has been given, i.e., $I \cup J = \{1,2,\cdots,m\}$ with $I \cap J = \varnothing$, and that the set K is given by

$$K = \{(z_1,\cdots,z_m) : i \in I \Rightarrow z_i \ge 0,\ i \in J \Rightarrow z_i \le 0\}.$$

The interior of K, K^0, will be defined to be

$$K^0 = \{(z_1,\cdots,z_m) : i \in I \Rightarrow z_i > 0,\ i \in J \Rightarrow z_i < 0\}.$$

Also throughout this paper it is assumed that there is one point on $[a,b]$ at which no row of $P(t)$ is zero, and that $P(t)$ and $R^{-1}(t)$ satisfy the following positivity condition:

$$\text{For every } t \in [a,b], \quad P(t): K \to K \quad \text{and} \quad R^{-1}(t): K \to K.$$

If a matrix A satisfies the above positivity condition, then A shall be said to be "positive with respect to the cone K." If K is given by $K = \{(z_1,z_2) : z_1 \ge 0,\ z_2 \le 0,\}$ then two examples of such a matrix are

$$\begin{bmatrix} 2 & 0 \\ -1 & 1 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 2 & -2 \\ -1 & 1 \end{bmatrix}.$$

2. Focal points. A point $f(a) \in [a,b]$ is called a focal point of a relative to (1), provided there is a nontrivial solution $x(t)$ of (1) satisfying

$$(3) \quad x^{(i)}(a) = 0, \quad i = 0,1,\cdots,k-1, \qquad [rx^{(k)}]^{(i)}(f(a)) = 0, \quad i = 0,1,\cdots,n-k-1,$$

and there is no nontrivial solution $z(t)$ of (1) which satisfies

$$z^{(i)}(a) = 0, \quad i = 0,1,\cdots,k-1, \qquad [rz^{(k)}]^{(i)}(c) = 0, \quad i = 0,1,\cdots,n-k-1,$$

for $a < c < f(a)$. If (1) does not possess such a focal point on $(a,b]$, the equation will be said to be disfocal on $[a,b]$.

For any matrix $A(t)\in C^{n-k}[a,b]$ for which $A^{-1}(t)$ exists on $[a,b]$, define the differential operator D by

$$D(u)=(-1)^{n-k}\left[A(t)u^{(k)}(t)\right]^{(n-k)},$$

$$u^{(i)}(a)=0,\quad i=0,1,\cdots,k-1,\qquad [Au^{(k)}]^{(i)}(b)=0,\quad i=0,1,\cdots,n-k-1.$$

Then the Green's function $g(t,s,a,A^{-1})$ is given by $g=g(t,s,a,A^{-1})$, where

$$g=\frac{1}{(n-k-1)!(k-1)!}\int_a^t (t-\xi)^{k-1}(s-\xi)^{n-k-1}A^{-1}(\xi)\,d\xi,\qquad a\le t\le s\le b,$$

and

$$g=\frac{1}{(n-k-1)!(k-1)!}\int_a^s (t-\xi)^{k-1}(s-\xi)^{n-k-1}A^{-1}(\xi)\,d,\qquad a\le s\le t\le b.$$

Then $x(t)$ satisfies the differential equation (1) and boundary conditions (3) if and only if $x(t)$ satisfies

$$x(t)=\int_a^b g(t,s,a,r^{-1})p(s)x(s)\,ds.$$

It will be useful to state and prove some lemmas at this time.

Let $\delta_i=1$ if $i\in I$ and $\delta_i=-1$ if $i\in J$.

LEMMA 1. *If $A=(A_{ij})$ is positive with respect to the cone K, then $\delta_i\delta_j A_{ij}\ge 0$, $i,j=1,\cdots,n$.*

Proof. Let e_j denote the jth unit basis vector. Then $\delta_j e_j\in K$ and $A(\delta_j e_j)\in K$. But the transpose of $A\delta_j e_j$ is $(\delta_j A_{ij},\cdots,\delta_j A_{mj})$, and this implies $\delta_i\delta_j A_{ij}\ge 0$.

LEMMA 2. *If $A=(A_{ij})$ is positive with respect to the cone K and $v\in K$, then $\delta_i A_{ij}v_j\ge 0$.*

Proof. Since $\delta_j v_j\ge 0$ and, from Lemma 1, $\delta_i\delta_j A_{ij}\ge 0$, it follows that $0\le\delta_j v_j\delta_i\delta_j A_{ij}=\delta_j A_{ij}v_j$.

LEMMA 3. *If A and B are both positive with respect to the cone K, then so is AB.*

LEMMA 4. *If A is positive with respect to K, then $A:K^0\to K^0$ if and only if no row of A is zero.*

LEMMA 5. *If $h(t)$ is continuous on $[a,b]$, $h(t):[a,b]\to K$ and there exists $t_0\in[a,b]$ such that $h(t_0)\in K^0$, then $\int_a^b h(s)\,ds\in K^0$.*

The main theorem can now be given.

THEOREM 1. *Suppose that $y(t)$ is a solution of (2) satisfying the boundary condition (3) and that $y(t)\in K^0$ for $t\in(a,b)$. If $\int_a^t|r_{ij}^{-1}(s)|\,ds\le\int_a^t|R_{ij}^{-1}(s)|\,ds$ and $|p_{ij}(t)|\le|P_{ij}(t)|$ for all $i,j\in(1,\cdots,m)$ and for all $t\in[a,b]$, and if furthermore for any $i=1,\cdots,m$, there exists an integer $j=j(i)$, $1\le j\le m$, and $t_i\in[a,b]$ such that $|p_{ij}(t_i)|<|P_{ij}(t_i)|$, then (1) is disfocal on $[\alpha,\beta]$, $[\alpha,\beta]\subset[a,b]$.*

Proof. Suppose, contrary to the conclusion of Theorem 1, that $x(t)$ is a nontrivial solution of (1) satisfying the boundary conditions

$$x^{(i)}(\alpha)=0,\quad i=0,1,\cdots,k-1,\qquad [rx^{(k)}]^{(i)}(\beta)=0,\quad i=0,1,\cdots,n-k-1,$$

for some $\alpha,\beta\in[a,b)$. Then of course,

$$x(t)=\int_\alpha^\beta g(t,s,\alpha,r^{-1})p(s)x(s)\,ds.$$

Also,

$$y(t)=\int_a^b g(t,s,a,R^{-1})P(s)y(s)\,ds.$$

Obviously $y^{(i)}(a)=0$, $i=0,1,\cdots,k-1$. It will now be shown that if $y_i(t)$ is the ith component of $y(t)$, then $y_i^{(k)}(a)\neq 0$. Toward this end it is easy to see by calculation that

$$y^{(k)}(t)=\frac{1}{(n-k-1)!}R^{-1}(t)\int_t^b (s-t)^{n-k-1}P(s)y(s)\,ds$$

and

$$y^{(k)}(a)=\frac{1}{(n-k-1)!}R^{-1}(a)\int_a^b (s-a)^{n-k-1}P(s)y(s)\,ds.$$

By hypothesis, $y(s)\in K^0$ for $s\in(a,b)$ and $P(s)y(s)\in K$ for $s\in(a,b)$. Also by hypothesis there exists a $t_0\in(a,b)$ such that no row of $P(t_0)$ is zero, and by Lemma 4 $P(t_0)y(t_0)\in K^0$. An application of Lemma 5 then shows that $\int_a^b(s-a)^{n-k-1}P(s)y(s)\,ds\in K^0$. Of course no row of $R^{-1}(a)$ can be zero, and using Lemma 4 again,

$$y^{(k)}(a)=\frac{1}{(n-k-1)!}R^{-1}(a)\int_a^b (s-a)^{n-k-1}P(s)y(s)\,ds\in K^0.$$

This then implies that no component of $y^{(k)}(a)$ can be zero. This shows that for any i, $y_i(t)$ has a zero at $t=a$ of precisely order k. Since $y(t)\in K^0$ for $t\in(a,b)$, no component of $y(t)$ can be zero, for $t\in(a,b)$. Thus for $\alpha\in[a,b)$, $y_i(t)$ has a zero of order at most k. Of course $x_i(t)$ has a zero at $t=\alpha$ of at least order k. Thus the terms $|x_i(t)|/|y_i(t)|$ are continuous on $(a,b]$ and, most importantly, are bounded on $(\alpha,b]$, for any $\alpha\geq a$. Define

$$\|x_i\|=\begin{cases}\sup\{|x_i(t)|/|y_i(t)|:t\in[\alpha,b]\} & \text{if } \alpha\in(a,b),\\ \sup\{|x_i(t)|/|y_i(t)|:t\in(\alpha,b]\} & \text{if } \alpha=a.\end{cases}$$

Also define $\|x\|=\max\{\|x_i\|:i=1,\cdots,m\}$.

For any $t\in(\alpha,b]$ if $\alpha=a$, or for any $t\in[\alpha,b]$ if $\alpha\in(a,b)$,

$$\begin{aligned}|x_\mu(t)|&=\left|\sum_{i,j}\int_\alpha^\beta g_{\mu i}(t,s,\alpha,r^{-1})p_{ij}(s)x_j(s)\,ds\right|\\ &\leq\sum_{i,j}\int_\alpha^\beta |g_{\mu i}(t,s,\alpha,r^{-1})|\,|p_{ij}(s)|\,|y_j(s)|\,|x_j(s)|\,|y_j^{-1}(s)|\,ds\\ &\leq\sum_{i,j}\int_\alpha^\beta |g_{\mu i}(t,s,\alpha,r^{-1})|\,|p_{ij}(s)|\,|y_j(s)|\,ds\|x\|.\end{aligned}$$

It follows readily from a calculation that $\partial/\partial\alpha$ of the last term is

$$-\frac{1}{(n-k-1)!(k-1)!}\cdot\sum_{i,j}|r_{\mu i}(\alpha)|\int_\alpha^\beta (t-\alpha)^{k-1}(s-\alpha)^{n-k-1}|p_{ij}(s)|\,|y_j(s)|\,ds\|x\|\leq 0.$$

Thus

$$|x_\mu(t)| \leq \sum_{i,j} \int_a^b |g_{\mu i}(t,s,a,r^{-1})|\,|p_{ij}(s)|\,|y_j(s)|\,ds\|x\|.$$

If $A(t)=(a_{ij}(t))$ is a matrix, define the $m\times m$ matrix A_1 by $(|a_{ij}(t)|)$. Then using the definition of $g(t,s,a,r^{-1})$, it follows immediately that $|g_{\mu i}(t,s,a,r^{-1})| \leq g_{\mu i}(t,s,a,r_1^{-1})$. Also notice that if $F(t,s,\xi)=(t-\xi)^{k-1}(s-\xi)^{n-k-1}$, then

$$\begin{aligned}
&(n-k-1)!(k-1)!g_{\mu i}(t,s,a,r_1^{-1}) \\
&\quad = \int_a^\delta (t-\xi)^{k-1}(s-\xi)^{n-k-1}|r_{ij}^{-1}(\xi)|\,d\xi \\
&\quad = \int_a^\delta F(t,s,\xi)|r_{ij}^{-1}(\xi)|\,d\xi \\
&\quad = \int_a^\delta |r_{ij}^{-1}(\xi)|\,d\xi F(t,s,\delta) - \int_a^\delta \int_a^\xi |r_{ij}^{-1}(\eta)|\,d\eta \frac{\partial F(t,s,\xi)}{\partial \xi}\,d\xi \\
&\quad \leq \int_a^\delta |R_{ij}^{-1}(\xi)|\,d\xi F(t,s,\delta) - \int_a^\delta \int_a^\xi |R_{ij}^{-1}(\eta)|\,d\eta \frac{\partial F(t,s,\xi)}{\partial \xi}\,d\xi \\
&\quad = \int_a^\delta (t-\xi)^{k-1}(s-\xi)^{n-k-1}|R_{ij}^{-1}(\xi)|\,d\xi,
\end{aligned}$$

since $F(t,s,\delta)\geq 0$ and $\partial F(t,s,\xi)/\partial\xi \leq 0$ over the regions of integrations. These remarks then show that

$$|x_\mu(t)| \leq \sum_{i,j} \int_a^b g_{\mu i}(t,s,a,R_1^{-1})|p_{ij}(s)|\,|y_j(s)|\,ds\|x\|. \tag{4}$$

But since $y_j(t)\neq 0$ on (a,b) for all j, and since by hypothesis $|p_{ij}(t_0)|<|P_{ij}(t_0)|$, and since at least one of the terms $R_{\mu i}^{-1}(t)$, $i=1,\cdots,m$, must not be zero for any $t\in[a,b]$, it can be seen that in the right-hand side of (4) the $|p_{ij}(s)|$ terms can be replaced with $|P_{ij}(s)|$ and obtain a strict inequality. That is,

$$|x_\mu(t)| < \sum_{i,j} \int_a^b g_{\mu i}(t,s,a,R_1^{-1})|P_{ij}(s)|\,|y_j(s)|\,ds\|x\|.$$

Since $g(t,s,a,R^{-1})$ and P are both positive with respect to the cone K, their product is also. Then by Lemma 2, the right-hand side of the last term is just

$$\delta_\mu \sum_{i,j} \int_a^b g_{\mu i}(t,s,a,R^{-1})P_{ij}(s)\,y_j(s)\,ds\|x\|,$$

where $\delta_i=1$ if $i\in I$ and $\delta_i=-1$ if $i\in J$. It then follows that

$$|x_\mu(t)|/|y_\mu(t)| = |x_\mu(t)|/\delta_\mu y_\mu(t),$$

and thus

$$\frac{|x_\mu(t)|}{|y_\mu(t)|} < \frac{1}{y_\mu(t)} \sum_{i,j} \int_a^b g_{\mu i}(t,s,a,R^{-1})P_{ij}(s)\,y_j(s)\,ds\|x\|. \tag{5}$$

If $\alpha \in (a,b)$, then (5) holds for $t \in [\alpha,b)$. Now assume that $\alpha = a$. It will be shown that this strict inequality also holds at $t \to a+$. This will be done by showing that the inequality

$$(6) \qquad \frac{1}{\delta_\mu y_\mu(t)} \sum_{i,j} \int_a^b g_{\mu i}(t,s,a,r_1^{-1}) |p_{ij}(s)| \, |y_j(s)| \, ds \|x\|$$

$$< \frac{1}{\delta_\mu y_\mu(t)} \sum_{i,j} \int_a^b g_{\mu i}(t,s,a,R_1^{-1}) |P_{ij}(s)| \, |y_j(s)| \, ds \|x\|,$$

which has already been established for $t \in (a,b)$, can be extended as a strict inequality to $[\alpha,b]$. To do this, we simply use l'Hôpital's rule and find that the limit as $t \to a+$ of the left-hand side of (6) is at most

$$(7) \qquad \frac{1}{\delta_i y_i^{(k)}(a)} \sum_{i,j} \frac{1}{(n-k-1)!} |R_{\mu i}^{-1}(a)| \int_a^b (s-t)^{n-k-1} |p_{ij}(s)| \, |y_j(s)| \, ds \|x\|,$$

and the limit as $t \to a+$ of the right-hand side of (6) is

$$(8) \qquad \frac{1}{\delta_i y_i^{(k)}(a)} \sum_{i,j} \frac{1}{(n-k-1)!} |R_{\mu i}^{-1}(a)| \int_a^b (s-t)^{n-k-1} |P_{ij}(s)| \, |y_j(s)| \, ds \|x\|.$$

Just as before, we readily see that indeed (7) is strictly less than (8). Since the left-hand side of (5) is less than or equal to the left-hand side of (6), and since as we have seen before the right-hand side of (5) equals the right-hand side of (6), inequality (5) is true on $[\alpha,b]$ even if $\alpha = a$.

Since $y(t)$ satisfies (2),

$$y_\mu(t) = \sum_{i,j} \int_a^b g_{\mu i}(t,s,a,R^{-1}) P_{ij}(s) \, y_j(s) \, ds$$

and the right-hand side of (5) is just $\|x\|$. This implies that $\|x_\mu\| < \|x\|$, and this in turn implies that $\|x\| < \|x\|$. From this contradiction, we infer the truth of the theorem.

The following theorem is more general than Theorem 1. The proof is the same as Theorem 1 with only the most obvious changes.

THEOREM 2. *Suppose that $y(t)$ is a solution of* (2) *satisfying the boundary conditions*

$$\left[R y^{(k)} \right]^{(i)}(b) = 0, \qquad i = 0, \cdots, n-k-1,$$

and that $y(t) \in K^0$ for all $t \in (a,b)$ and $y^{(i)}(a) \in K$ for $i = 0,1,\cdots,k-1$. If $\int_a^t |r_{ij}^{-1}(s)| ds \le \int_a^t |R_{ij}^{-1}(s)| ds$ and $|p_{ij}(t)| \le |P_{ij}(t)|$ for all $i,j \in (1,\cdots,m)$ and for all $t \in [a,b]$, and if furthermore for any $i = 1,\cdots,m$, there exists an integer $j = j(i)$, $1 \le j \le m$, and $t_i \in [a,b]$ such that $|p_{ij}(t)| < |P_{ij}(t_i)|$, then (1) *is disfocal on $[\alpha,\beta] \subset [a,b]$.*

Theorems 1 and 2 require $y(t)$ not only to be a solution of (2) satisfying the boundary condition (3) but also to satisfy $y(t) \in K^0$ for all $t \in (a,b)$. Using the theory of μ_0-positive linear operators, it can readily be shown that such solutions do indeed exist for a large class of matrices $R(t)$ and $P(t)$. Such a condition is to assume that $P(t)$ satisfies the additional assumption that for $v \in K$ and $v \neq 0$, $P(t)v \in K^0$ for all $t \in [a,b.]$. Notice that this latter condition, among other things, restricts $P(t)$ never to be diagonal, whereas $P(t)$ could be diagonal before. Notice also that the first example given at the beginning does not satisfy this latter condition, but that the second example does. The following theorem can now be given.

THEOREM 3. *Suppose that $P(t)$ also satisfies the condition that for any $v\in K$, $v\neq 0$ and $t\in[a,b]$, then $P(t)v\in K^0$. Then if* (2) *has a focal point b, then* (2) *has a solution $y(t)$ that satisfies the boundary conditions* (3) *and $y(t)\in K^0$ for $t\in(a,b)$.*

Proof. To establish this theorem define the Banach space

$$B=\{u\in C([a,b]):u(a)=0\},$$

with $\text{norm}\|u\|_1=\sup\{|u(t)|:t\in[a,b]\}$. Also define the cone

$$\tilde{K}=\{u\in B:u(t)\in K \text{ for } t\in[a,b]\},$$

with interior $\tilde{K}^0=\{u\in B:u(t)\in K^0 \text{ for } t\in(a,b)\}$. It is apparent that the operator

$$T(u)=\int_a^b g(t,s,a,R^{-1})P(s)u(s)\,ds$$

maps $\tilde{K}$ into $\tilde{K}^0$, and therefore by arguments similar to that found in [2], T is μ_0-positive with respect to the cone $\tilde{K}$. From this, Theorem 3 follows in the same way as found in [2].

We can now state a comparison theorem for focal points of (1) and (2). Let $f_p(a)$ and $f_P(a)$ be the focal points of (1) and (2), respectively.

THEOREM 4. *Suppose that $P(t)$ satisfies the additional condition that for every $v\in K$, $v\neq 0$, $t\in[a,b]$, $P(t)v\in K^0$. If $\int_a^t|r_{ij}^{-1}(s)|ds\leq\int_a^t|R_{ij}^{-1}(s)|ds$ and if $|p_{ij}(t)|\leq|P_{ij}(t)|$ for $i,j\in(1,\cdots,m)$ and for all $t\in[a,f_p(a)]$, and if furthermore for any $i=1,\cdots,m$, there exists an integer $j=j(i)$, $1\leq j\leq m$, and $t_i\in[a,f_p(a)]$ such that $|p_{ij}(t_i)|<|P_{ij}(t_i)|$, then $f_p(a)>f_p(a)$.*

Proof. This theorem follows immediately from Theorems 1 and 3.

In all the theorems presented here, it has always been assumed that the absolute value of $p_{ij}(t)$ must be less than or equal to $|P_{ij}(t)|$, in order to assure that (1) oscillates slower than (2). This condition obviously says that $p_{ij}(t)$ cannot be too negative. The following example will demonstrate that the theorems presented here are not true if one just assumes that $p_{ij}(t)\leq|P_{ij}(t)|$. Let P be the identity matrix and let the symmetric matrix p be given by

$$p=\begin{bmatrix}1 & -1\\ -1 & 1\end{bmatrix}$$

and $n=2$, $k=1$. Of course $f_P(0)=\pi/2$. Notice that a solution of $x''+px=0$ is $x=(\sin\sqrt{2}\,t,\ -\sin\sqrt{2}\,t)/\sqrt{2}$, and it is easy to see that $f_p(0)=(\pi/2)\sqrt{2}<f_P(0)$. This example clearly shows that $p<P$ (in the sense that $p_{ij}(t)\leq P_{ij}(t)$) does not imply $f_P(a)<f_p(a)$, even in the selfadjoint case. This example also illustrates just how useful Theorem 1 can be. Change the roles of p and P, i.e., define p to be the identity matrix and P to be the other matrix. Notice that P is positive with respect to the cone $K=\{(z_1,z_2):z_1\geq 0,z_2\leq 0\}$ and that $y(t)=(\sin\sqrt{2}\,t,-\sin\sqrt{2}\,t)/\sqrt{2}$ is a solution of $y''+P(t)y=0$, $y(0)=0$, $y'(\pi/2\sqrt{2})=0$ and $y(t)\in K^0$ for $t\in(0,\pi/2\sqrt{2})$. Also $|p_{ij}(t)|\leq|P_{ij}(t)|$ and $|p_{12}|<|P_{12}|$. Thus Theorem 1 implies that $x''+p(t)x=0$ is disfocal on $[0,\pi/2\sqrt{2}]$, which it is, despite the fact that in some sense p is larger than P. By the way, it is interesting to notice that in these last examples that neither $p-P$ nor $P-p$ is positive semidefinite and that thus the classical theorems cannot apply.

REFERENCES

[1] U. ELIAS, *Focal points for a linear differential equation whose coefficients are of constant signs*, Trans. Amer. Math. Soc., 249 (1979), pp. 187–202.

[2] R. GENTRY AND C. TRAVIS, *Comparison of eigenvalues associated with linear differential equations of arbitrary order*, Trans. Amer. Math. Soc., 223 (1976), pp. 167–179.

[3] M. KEENER AND C. TRAVIS, *Focal points and positive cones for a class of n-th order differential equations*, Trans. Amer. Math. Soc., 237 (1978), pp. 331–351.

[4] _____, *Sturmian theory for a class of nonselfadjoint differential systems*, Ann. Mat. Pura Appl., CXXIII (1980), pp. 247–266.

[5] M. MORSE, *A generalization of the Sturm separation and comparison theorems in n-space*, Math. Ann., 103 (1930), pp. 53–69.

[6] Z. NEHARI, *Oscillation criteria for second order linear differential systems*, Trans. Amer. Math. Soc., 85 (1957), pp. 428–445.

[7] _____, *Green's functions and disconjugacy*, Arch. Rat. Mech. Anal., 62 (1976), pp. 53–67.

[8] W. T. REID, *Ordinary Differential Equations*, John Wiley, New York, 1971.

[9] E. TOMASTIK, *Comparison theorems for second order nonselfadjoint differential systems*, this Journal, 14 (1983), pp. 60–65.

[10] C. TRAVIS, *Comparison of eigenvalues for linear differential equations of order* $2n$, Trans. Amer. Math. Soc., 177 (1973), pp. 363–374.

SIAM J. MATH. ANAL.
Vol. 17, No. 1, January 1986

005

ON THE STABILITY OF A DISTRIBUTED NETWORK*

JOSÉ M. FERREIRA†

Abstract. This paper is devoted to the study of the stability of a lossless transmission line network which is described by a one delay differential-difference equation of neutral type, involving three parameters. It is shown that for certain intervals of these parameters, asymptotic stability is available when the delay verifies a boundedness condition.

1. Introduction. This paper is devoted to the study of the asymptotic stability of the differential-difference equation of neutral type

$$\dot{x}(t)+k\dot{x}(t-r)+\alpha x(t)-\alpha k x(t-r)-h(x(t))-kh(x(t-r))=0, \tag{1.1}$$

where $r>0$, $k\neq 0$, $\alpha>0$ and $h\in C^1(\mathbb{R})$ is such that $h(0)=0$. According to [1], [6], this equation arises from a lossless transmission line network.

The exponential asymptotic stability of the zero solution of (1.1) is determined by the exponential asymptotic stability of the linearized problem, which in turn occurs if and only if for some $\varepsilon>0$, all roots of the characteristic equation

$$z(1+k\exp(-rz))+a-kb\exp(-rz)=0 \tag{1.2}$$

satisfy $\operatorname{Re} z\leqq -\varepsilon$, where $\gamma=h'(0)$, $a=\alpha-\gamma$ and $b=\alpha+\gamma$.

The case $\gamma<\alpha$ is considered in [1], [2], [6] (see also [3]). In [1], [2] in working on the characteristic equation, the exponential asymptotic stability of the linearized problem is shown for $\gamma<\alpha(1-|k|)/(1+|k|)$, as well as the existence of oscillations for certain choices of k and $\gamma\in]\alpha(1-|k|)/(1+|k|),\alpha[$. By use of Lyapunov functionals, Slemrod [6] shows the uniform asymptotic stability of (1.1) if $\sup_{x>0}(h(x)/x)<\alpha(1-|k|)/(1+|k|)$.

We will discuss here the case $\gamma\geqq\alpha$, that is when $0\leqq -a<b$. With respect to the parameter γ, the interval $[\alpha,+\infty[$ is given as a potential region of instability. This will be confirmed in §§2, 3. However imposing some bounds to the delay r, asymptotic stability is still available for some negative values of k. This is shown in §3.

Notice that when $k=0$, (1.2) has a solution with a nonnegative real part ($z=-a\geqq 0$), and as we will see this situation is maintained for $a/b\leqq k<0$. Therefore we cannot apply an argument of continuity of the spectra, as in [1], [2], in order to conclude asymptotic stability. We will prove that for k in some interval $]\beta,a/b[$, asymptotic stability occurs and disappears when $-1<k\leqq\beta$. This means that part of the spectrum of the linearized problem moves from the right-half complex plane to the left one and returns to the first after a certain value of k. This is a kind of situation which does not happen in the case $\gamma<\alpha$, where the whole spectrum is moving from the left-half complex plane to the right.

The linearized problem of (1.1) is exponentially asymptotically stable only if all solutions of the associated difference equation

$$x(t)+kx(t-r)=0 \tag{1.3}$$

*Received by the editors July 7, 1982, and in revised form April 2, 1984.
†Centro de Fisica da Materia Condensada, Avenida Prof. Gama Pinto, 2, 1699 Lisboa Codex, Portugal.

exhibit an exponential decay [4]. In order to obtain this decay we will always assume $|k|<1$, since then all roots of

$$1+k\exp(-rz)=0 \tag{1.4}$$

satisfy $\operatorname{Re} z=(1/r)\log|k|<0$. As a preliminary result we have:

THEOREM 1.1. *The real part of all roots of* (1.2) *is less than* $c=(b-a)/2>0$. *For any* $0<\varepsilon<\delta(k,r)=-(1/r)\log|k|$, *the set of all roots of* (1.2), *such that* $\operatorname{Re} z\geqq-\varepsilon$, *is bounded.*

Proof. Assume that $z=u+iv$ is a root of (1.2). Since (1.2) can be written as

$$(z+a)+k(z-b)\exp(-rz)=0, \tag{1.5}$$

we have

$$\exp(ru)=\frac{|k||z-b|}{|z+a|}. \tag{1.6}$$

If $u\geqq c$ then $|z-b|\leqq|z+a|$, and consequently we obtain $u=\{\log(|k||z-b|/|z+a|)\}/r<0$, which is contradictory.

For the second part of the theorem, taking $M=1-|k|\exp(r\varepsilon)$ and z in the strip $-\varepsilon\leqq\operatorname{Re} z\leqq c$, we have $|1+k\exp(-rz)|\geqq M$. Thus if z is a root of (1.2) in the mentioned strip, we have $|z|\leqq(|a|+|b|)/M$. The statement then follows.

By analyticity, if all roots of (1.2) stay in the left-half complex plane, then by Theorem 1.1 they satisfy $\operatorname{Re} z\leqq-\varepsilon$ for some $0<\varepsilon<\delta(k,r)$. Therefore if we consider the rectangle $R_\nu=\{z:\ 0\leqq\operatorname{Re} z\leqq c,|\operatorname{Im} z|\leqq\nu\}$ for $\nu>0$, (1.2) has all roots satisfying $\operatorname{Re} z\leqq-\varepsilon$ for some $\varepsilon>0$ if and only if the function

$$f(z)=g(z)+k\exp(-rz),$$

where $g(z)=(z+a)/(z-b)$, has no roots in R_ν for every $\nu>0$.

Considering the boundary Γ_ν of R_ν, by the argument's principle, the existence and nonexistence of zeros with a nonnegative real part depends upon the variation of $f(z)$ when z proceeds along Γ_ν. In order to study this variation we develop a method introduced in [5], which consists in comparing the variation of $g(z)$ to the variation of $k\exp(-rz)$.

Making $z=u+iv$, we will write $g(z)$ as

$$g(u+iv)=G(u+iv)\exp(i\theta(u+iv)), \tag{1.7}$$

where

$$G(u+iv)=\frac{\left[(u+a)^2+v^2\right]^{1/2}}{\left[(u-b)^2+v^2\right]^{1/2}} \tag{1.8}$$

and

$$\theta(u+iv)=\arctan\left\{\frac{-(b+a)v}{\left[(u+a)(u-b)+v^2\right]}\right\}. \tag{1.9}$$

2. Oscillations. An oscillation for the linearized problem appears whenever $f(z)$ has a zero on the imaginary axis. Since $f(\bar z)=\overline{f(z)}$, it will be sufficient to study $f(z)$ for z along the nonnegative imaginary half-axis $\{z:\ z=iv, v\geqq0\}$. The same holds for $g(z)$, which study is stated in the following.

LEMMA 2.1. *For* $v \geqq 0$, $\theta(iv)$ *decreases in* $[0,(-ab)^{1/2}]$ *and increases in* $[(-ab)^{1/2}, +\infty)$; $\theta(0)=0$ *and* $-\pi/2<\theta(iv)<0$, $v\neq 0$. *On the other hand,* $G(iv)$ *increases in* $v\geqq 0$, $G(iv)<1$ *and* $\lim_{v\to+\infty} G(iv)=1$.

Proof. Make $u=0$ in (1.9) and (1.8). The sign of $(d/dv)\theta(iv)$ depends upon the sign of $ab+v^2$. In fact $\theta(iv)$ decreases as long as $ab+v^2\leqq 0$ and increases when $ab+v^2\geqq 0$.

As $g(0)>0$, then $\theta(0)=0$. Since $-\infty<\tan\theta(iv)<0$ for $v\neq 0$, we have $-\pi/2<\theta(iv)<0$. Moroever, $\lim_{v\to+\infty}\theta(iv)=0$.

On the other hand, as $b>-a$ then $G(iv)<1$ and $\lim_{v\to+\infty} G(iv)=1$. By derivation one easily sees that $G(iv)$ is increasing. □

Then, assuming $G(0)\leqq|k|$, let $v_0\geqq 0$ be such that $G(iv_0)=|k|$. In order to have $f(iv)=0$, it is necessary and sufficient that $v=v_0$ and that the origin be in the line segment of the complex plane

$$\langle g(iv), k\exp(-irv)\rangle = \{(1-\lambda)g(iv)+\lambda k\exp(-irv);\ 0\leqq\lambda\leqq 1\}.$$

The origin is aligned with $g(iv)$ and $k\exp(-irv)$ whenever

$$\arg g(iv)=\arg(k\exp(-irv))+n\pi \tag{2.1}$$

for some integer n. For $n=0$, $v=0$, we have always an initial alignment in the real axis.

If $k<0$, condition (2.1) becomes

$$\theta(iv)+rv=(n+1)\pi,$$

and since $-\pi/2<\theta(iv)\leqq 0$, we have that $n\geqq -1$. However, the alignments which have the origin in $\langle g(iv), k\exp(-irv)\rangle$ are given by (see Fig. 1)

$$\theta(iv)+rv=2n\pi \quad\text{and}\quad \frac{2n\pi}{r}\leqq v<\frac{(2n+1/2)\pi}{r}, \qquad n\geqq 0. \tag{2.2}$$

When $n\geqq 1$, as for $rv=2n\pi$, we have $\theta(iv)+rv<2n\pi$, and for $rv=(2n+1/2)\pi$, $\theta(iv)+rv>2n\pi$. By continuity we conclude that there is always an $\omega_n\in]2n\pi/r, (2n+1/2)\pi/r[$, which verifies (2.2). For $n=0$, if there exists an $\omega_0\in]0, \pi/2r[$ such that $\theta(i\omega_0)=-r\omega_0$, then we obtain another alignment, and $0\in\langle g(i\omega_0), k\exp(-ir\omega_0)\rangle$. This will happen if and only if $(d/dv)\theta(iv)|_{v=0}<-r$; that is, $r<(b+a)/(-ab)$.

If $k>0$, condition (2.1) can be written as

$$\theta(iv)+rv=n\pi,$$

where $n\geqq 0$. The alignments which have the origin in $\langle g(iv), k\exp(-irv)\rangle$ are given by $\theta(iv)+rv=(2n-1)\pi$ and $(2n-1)\pi/r<v<(2n-1/2)\pi/r$, $n\geqq 1$. As above, this condition is always verified for some ω_n in $](2n-1)\pi/r, (2n-1/2)\pi/r[$. Hence, $f(iv)=0$ if and only if $v=v_0=\omega_n$ for some $n\geqq 1$.

Notice that when $(d/dv)\theta(iv)<0$, we have $(d^2/dv^2)\theta(iv)\geqq 0$. This means that for $v\in[0,(-ab)]$, $\theta(iv)$ is a convex function. Therefore all the alignments mentioned above are unique in each considered interval.

The conditions obtained for the existence of oscillations are equivalent to those of Brayton in [1].

3. Stability for bounded delays. When we compare $g(z)$ with $k\exp(-rz)$, different conclusions may be obtained according to the signal of k. In the following we show that for $k<0$, asymptotic stability can be obtained.

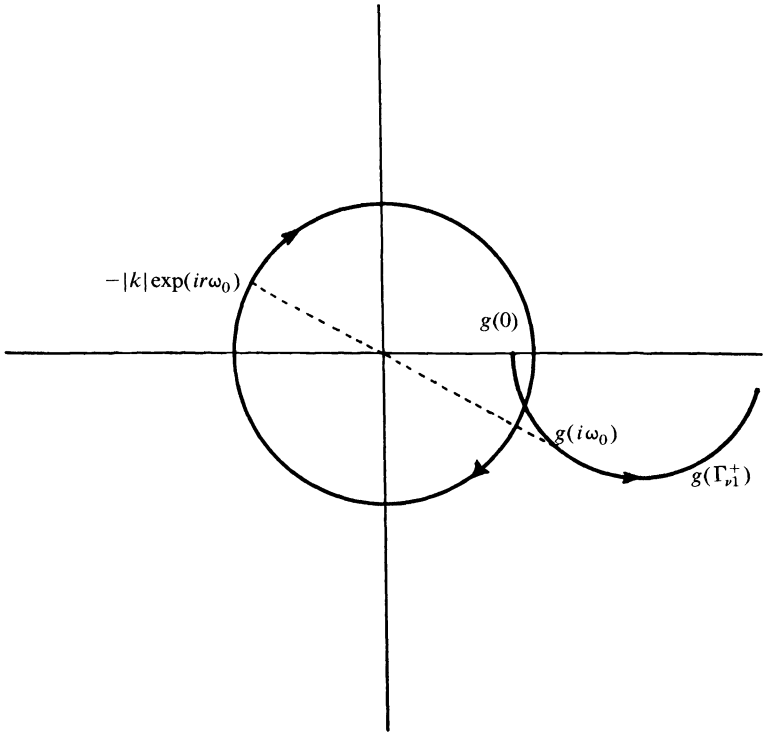


FIG. 1

When $-1<k<0$ we write $f(z)$ as

$$f(z)=g(z)-|k|\exp(-rz).$$

Therefore if $g(0)=|k|$ we have $f(0)=0$. We avoid this degenerate situation by assuming $g(0)\neq|k|$.

We have already seen that when $r<(b+a)/(-ab)$, there exists an $\omega_0\in]0, \pi/2r[$ such that $0\in\langle g(i\omega_0), -|k|\exp(-ir\omega_0)\rangle$. We will show:

THEOREM 3.1. *Let* $-1<k<a/b$ *and* $r<-(b+a)/(ab)$. *If* $|g(i\omega_0)|>|k|$, *then all zeros of* $f(z)$ *have negative real part. If* $|g(i\omega_0)|<|k|$, $f(z)$ *has two zeros with positive real part such that* $0<|\mathrm{Im}\, z|<\pi/r$. *If* $|g(i\omega_0)|=|k|$, *then* $f(\pm i\omega_0)=0$.

In order to prove this theorem we will consider the nonnegative imaginary part of the rectangle R_ν,

$$R_\nu^+=\{z: 0\leqq \mathrm{Re}\, z\leqq c, 0<\mathrm{Im}\, z\leqq \nu\},$$

which boundary we denote by Γ_ν^+. Decompose

$$\Gamma_\nu^+=\Gamma_{\nu 1}^+\cup\Gamma_{\nu 2}^+\cup\Gamma_{\nu 3}^+\cup\Gamma_{\nu 0},$$

where

$$\Gamma_{\nu 1}^+=\{iv: 0\leqq v\leqq \nu\},\qquad \Gamma_{\nu 2}^+=\{u+i\nu: 0\leqq u\leqq c\},$$

$$\Gamma_{\nu 3}^+=\{c+iv: 0\leqq v\leqq \nu\},\quad \Gamma_{\nu 0}=\{u: 0\leqq u\leqq c\}.$$

As we want to compare $g(z)$ with $-|k|\exp(-rz)$ when z proceeds along Γ_ν^+ clockwise from the origin, we first need to know the variation of $g(z)$ along Γ_ν^+. We recall that for $z=iv$ in $\Gamma_{\nu 1}^+$, $\theta(iv)$ is always decreasing if $ab+\nu^2\leqq 0$. *If* $ab+\nu^2>0$, $\theta(iv)$ decreases in $[0,(-\mathrm{ab})^{1/2}]$ and increases in $[(-ab)^{1/2},\nu]$. Moreover, $G(iv)$ is increasing and is less than one.

LEMMA 3.1. *For* $z=u+iv$ *along* $\Gamma_{\nu 2}^+$, $\theta(u+iv)$ *decreases. If* $(b+a)^2\leqq 4\nu^2$, *it is* $-\pi/2\leqq\theta(u+iv)<0$ *for* u *in* $[0,c]$; $\theta(u+iv)=-\pi/2$ *is satisfied only when* $(b+a)^2=4\nu^2$ *and* $u=c$. *If* $(b+a)^2>4\nu^2$, *there exists a* $u_0\in\,]0,c[$ *such that* $-\pi/2<\theta(u+iv)<0$ *in* $[0,u_0[$; $-\pi<\theta(u+iv)<-\pi/2$ *in* $]u_0,c]$ *and* $\theta(u_0+iv)=-\pi/2$. *Moreover,* $G(u+iv)<1$ *for* $u<c$ *and* $G(c+iv)=1$. *If* $\nu^2+ab\geqq 0$, $G(u+iv)$ *increases. If* $\nu^2+ab<0$, *there exists a* $u_1\in\,]0,c[$ *such that* $G(u+iv)$ *decreases in* $[0,u_1]$ *and increases in* $[u_1,c]$.

Proof. We have $\theta(u+iv)=\arctan\{-(b+a)\nu/[(u+a)(u-b)+\nu^2]\}$ and $(d/du)\theta(u+iv)<0$, except when $u=c$, where this derivative is zero. Then $\theta(u+iv)$ decreases. If $(b+a)^2<4\nu^2$, we have $p(u)=(u+a)(u-b)+\nu^2>0$ for every $u\geqq 0$, and then $-\pi/2<\theta(u+iv)<0$. For $(b+a)^2=4\nu^2$ the same holds except when $p(u)=0$, which happens only for $u=c$. Only in this case does $\theta(u+iv)=-\pi/2$. If $(b+a)^2>4\nu^2$, then $p(u)$ has only one positive root u_0 in $]0,c[$. We have $\theta(u_0+iv)=-\pi/2$, $\theta(u+iv)>-\pi/2$ for $0\leqq u<u_0$ and $\theta(u+iv)<-\pi/2$ for $u_0<u\leqq c$.

On the other hand, clearly $G(u+iv)\leqq 1$ and $G(u+iv)=1$ only if $u=c$. Moreover, as $(d/du)G(u+iv)^2=2(b+a)[\nu^2-(u+a)(u-b)]/[(u-b)^2+\nu^2]^2$, $G(u+iv)$ increases if $\nu^2+ab\geqq 0$. If $\nu^2+ab<0$, the polynomial $\nu^2-(u+a)(u-b)$ has only one positive real zero $u_1\in\,]0,c[$ and $G(u+iv)$ decreases in $[0,u_1]$ and increases in $[u_1,c]$. □

LEMMA 3.2. *For* $z=c+iv$ *moving along* $\Gamma_{\nu 3}^+$, *when* v *goes from* ν *to zero,* $\theta(c+iv)$ *decreases to* $\theta(c)=-\pi$. *If* $(b+a)>2\nu$, *it is* $-\pi\leqq\theta(c+iv)<-\pi/2$. *If* $(b+a)=2\nu$, *the same holds except for* $\theta(c+iv)=-\pi/2$. *If* $(b+a)<2\nu$, *there exists a* $v_1\in\,]0,\nu[$ *such that* $\theta(c+iv_1)=-\pi/2$, $\theta(c+iv)>-\pi/2$ *for* $v>v_1$ *and* $\theta(c+iv)<-\pi/2$ *for* $v<v_1$. *Furthermore,* $G(c+iv)=1$ *for every* $v\in[0,\nu]$.

Proof. When $z=c+iv$ we have $(d/dv)\theta(c+iv)>0$. Therefore when v goes from ν to zero, $\theta(c+iv)$ decreases. As $g(c)=-1$, it is $\theta(c)=-\pi$. One can easily see that $G(c+iv)=1$ for every v. The rest of the lemma follows as in Lemma 3.1. □

Let $g(0)=-a/b>|k|$ and recall the boundary Γ_ν of the rectangle R_ν. Taking ν sufficiently large, one can see by Lemmas 3.1 and 3.2 that $|g(z)|>|k|$ for every z in Γ_ν. Therefore as $|f(z)-g(z)|\leqq|k|$ for z in Γ_ν, we conclude by Rouché's theorem that f and g have the same number of zeros inside Γ_ν. Hence if $g(0)>|k|$, f has a unique zero, which is necessarily real, inside Γ_ν. This conclusion, which can also be reached in a more direct way, is the reason why in Theorem 3.1 we have to assume $g(0)=-a/b<|k|$.

Assume now $G(i\omega_0)>|k|$ and take $\nu=m\pi/r$, where m is a positive even integer arbitrarily large. Notice that then $v_0<\omega_0$.

For $z=u+iv$, the real and imaginary parts of $f(z)$ are

$$\begin{aligned}\mathrm{Re}f(u+iv)&=G(u+iv)\cos\theta(u+iv)-|k|\exp(-ru)\cos rv,\\ \mathrm{Im}f(u+iv)&=G(u+iv)\sin\theta(u+iv)+|k|\exp(-ru)\sin rv.\end{aligned}\tag{3.1}$$

(A) We will analyze here the variation of $f(z)$ for $z=iv$ along $\Gamma_{\nu 1}^+$.

(1) Let $v\in[0,\pi/2r]$. (i) for $v=0$ we have $\mathrm{Im}f(0)=0$ and $\mathrm{Re}f(0)=G(0)-|k|<0$. (ii) For $0<v\leqq v_0$, it is $-\pi/2<\theta(iv)<-rv\leqq 0$. Then $\cos\theta(iv)<\cos rv$ and $\mathrm{Re}f(iv)<(G(iv)-|k|)\cos rv\leqq 0$. (iii) When $v_0\leqq v\leqq\omega_0$, we have $-\pi/2<\theta(iv)\leqq -rv\leqq 0$ and $\sin\theta(iv)\leqq-\sin rv$. Thus, for $v<\omega_0$, $\mathrm{Im}f(iv)<(-G(iv)+|k|)\sin rv\leqq 0$. Moreover,

$\mathrm{Im} f(i\omega_0)=(-G(i\omega_0)+|k|)\sin r\omega_0<0$. (iv) If $\omega_0\leqq v\leqq \pi/2r$ then $\theta(iv)\geqq -rv$, which implies $\mathrm{Re} f(iv)>(G(iv)-|k|)\cos rv\geqq 0$ for $v>\omega_0$. Furthermore, $\mathrm{Re} f(i\omega_0)>0$.

(2) In each interval $[(2n+1/2)\pi/r,(2n+1)\pi/r]$, $n\geqq 0$, we have always $\mathrm{Re} f(iv)>0$.

(3) If $v\in[(2n+1)\pi/r,(2n+2)\pi/r]$, $n\geqq 0$ then $\mathrm{Im} f(iv)<0$.

(4) Now let $v\in[2n\pi/r,(2n+1/2)\pi/r]$ for $n\geqq 1$, and take the alignment ω_n in the interior of this interval. (i) For $2n\pi/r\leqq v\leqq\omega_n$ we have $\theta(iv)\leqq -rv+2n\pi$. Thus, as in (1) (iii), it is $\mathrm{Im} f(iv)<0$. (ii) When $\omega_n\leqq v\leqq(2n+1/2)\pi/r$ as in (1) (iv), we have $\mathrm{Re} f(iv)>0$.

Thus, for $z=iv$ moving along $\Gamma_{\nu 1}^{+}$, $f(z)$ starts in the left-half plane and passes into the right one at some v in $[v_0,\omega_0]$ with $\mathrm{Im} f(z)<0$. After that we have always either $\mathrm{Re} f(z)>0$ or $\mathrm{Im} f(z)<0$ (see Fig. 2 below).

(B) When $z=u+i\nu$ proceeds along $\Gamma_{\nu 2}^{+}$, we conclude by Lemma 3.1 that

$$\mathrm{Im} f(u+i\nu)=G(u+i\nu)\sin\theta(u+i\nu)<0.$$

(C) Now we will discuss the variation of $f(z)$ when $z=c+iv$ moves along $\Gamma_{\nu 3}^{+}$. By Lemma 3.2, for every $v\in[(2n+1)\pi/r,(2n+2)\pi/r]$, $n\geqq 0$, we have $\mathrm{Im} f(c+iv)=G(c+iv)\sin\theta(c+iv)+|k|\exp(-rc)\sin rv<0$. In each interval $[2n\pi/r,(2n+1)\pi/r]$, $n\geqq 0$, alignments exist for $g(c+iv)$, $-|k|\exp(-r(c+iv))$ and the origin such that $0\in\langle g(c+iv),\ -|k|\exp(-r(c+iv))\rangle$. In fact, as in §2, for each $n\geqq 0$, there exists a $\tau_n\in]2n\pi/r,(2n+1)\pi/r[$ such that $\theta(c+i\tau_n)+r\tau_n=2n\pi$. A final alignment exists along the negative real axis, when $n=0$, $v=0$—that is, for $z=c$.

(1) Let $\tau_n=(2n+1/2)\pi/r$. (i) Then $\theta(c+i\tau_n)=-\pi/2$, and consequently $\mathrm{Re} f(c+i\tau_n)=0$ and $\mathrm{Im} f(c+i\tau_n)=-1+|k|\exp(-rc)<0$. (ii) For $2n\pi/r\leqq v<(2n+1/2)\pi/r$ we have $\theta(c+iv)<\theta(c+i\tau_n)=-\pi/2$, and then $\mathrm{Re} f(c+iv)=\cos\theta(c+iv)-|k|\exp(-rc)\cos rv<0$. (iii) When $(2n+1/2)\pi/r<v\leqq(2n+1)\pi/r$ it is $\theta(c+iv)>-\pi/2$, and therefore $\mathrm{Re} f(c+iv)>0$.

(2) Now assume $\tau_n>(2n+1/2)\pi/r$. Then $\theta(c+i\tau_n)<-\pi/2$. (i) For $2n\pi/r\leqq v\leqq(2n+1/2)\pi/r$ we have $\theta(c+iv)<-\pi/2$, which implies $\mathrm{Re} f(c+iv)<0$. (ii) For $(2n+1/2)\pi/r\leqq v\leqq\tau_n$ we have $-\pi<\theta(c+iv)\leqq -rv+2n\pi\leqq -\pi/2$. Therefore $\cos\theta(c+iv)\leqq\cos(-rv+2n\pi)=\cos rv$ and $\mathrm{Re} f(c+iv)<(1-|k|\exp(-rc))\cos rv\leqq 0$ if $v<\tau_n$. On the other hand, $\mathrm{Re} f(c+i\tau_n)=(1-|k|\exp(-rc))\cos r\tau_n<0$. (iii) For $\tau_n\leqq v\leqq(2n+1)\pi/r$ and $\theta(c+iv)<-\pi/2$ we have $-\pi/2>\theta(c+iv)\geqq -rv+2n\pi>-\pi$, and then $\sin\theta(c+iv)\leqq-\sin rv$, which implies $\mathrm{Im} f(c+iv)<(-1+|k|\exp(-rc))\sin rv\leqq 0$ if $v>\tau_n$. On the other hand, $\mathrm{Im} f(c+i\tau_n)<0$. (iv) If $\tau_n\leqq v\leqq(2n+1)\pi/r$ and $\theta(c+iv)\geqq-\pi/2$, we have $\mathrm{Re} f(c+iv)>0$.

(3) Let $\tau_n<(2n+1/2)\pi/r$. Then $\theta(c+i\tau_n)>-\pi/2$. (i) For $2n\pi/r\leqq v\leqq\tau_n$ and $\theta(c+iv)\leqq-\pi/2$ we obtain $\mathrm{Re} f(c+iv)<0$. (ii) When $2n\pi/r\leqq v\leqq\tau_n$ and $\theta(c+iv)\geqq-\pi/2$, it is $-\pi/2\leqq\theta(c+iv)\leqq -rv+2n\pi\leqq 0$. Thus $\sin\theta(c+iv)\leqq-\sin rv$, and consequently $\mathrm{Im} f(c+iv)<(-1+|k|\exp(-rc))\sin rv\leqq 0$ if $v<\tau_n$. Clearly $\mathrm{Im} f(c+i\tau_n)<0$. (iii) When $\tau_n\leqq v\leqq(2n+1/2)\pi/r$, we have $0>\theta(c+iv)\geqq -rv+2n\pi\geqq-\pi/2$. Therefore $\mathrm{Re} f(c+iv)>0$. (iv) For $(2n+1/2)\pi/r\leqq v\leqq(2n+1)\pi/r$ it is $\theta(c+iv)>-\pi/2$, and then $\mathrm{Re} f(c+iv)>0$.

Recalling that $\theta(c)=-\pi$, for $z=c+iv$ moving along $\Gamma_{\nu 3}^{+}$, $f(z)$ goes back to the left-half plane, but whenever $f(z)$ crosses the imaginary axis it does with $\mathrm{Im} f(z)<0$.

(D) Finally when $z=u$ moves along $\Gamma_{\nu 0}$, we have $f(u)=(u+a)/(u-b)-|k|\exp(-ru)$, and since $r<-(b+a)/(ab)$ and $-a/b<|k|$, one can easily prove that $f(u)$ decreases when u increases in $[0,c]$. Therefore $f(z)$ has no zero on $\Gamma_{\nu 0}$, and when z moves along $\Gamma_{\nu 0}$ from c to zero, $f(z)$ moves along the negative real axis from $f(c)=-1-|k|\exp(-rc)$ to $f(0)=-a/b-|k|$.

Thus by the analysis made above, when z proceeds along Γ_ν^+ clockwise, $f(z)$ passes from the left- to the right-half complex plane and returns to the first. Whenever $f(z)$ crosses the imaginary axis we have $\mathrm{Im} f(z)<0$. This means that $f(\Gamma_\nu^+)$ never encircles the origin and that then $f(z)$ has no zeros in R_ν^+. Hence all zeros of $f(z)$ lie in the left-half plane.

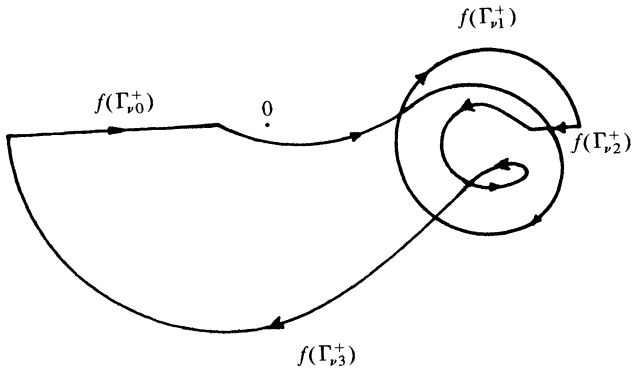


FIG. 2

Assume now $G(i\omega_0)<|k|$ and take $\nu=\pi/r$. It will be $v_0>\omega_0$, and we distinguish two cases: $v_0\leqq\pi/2r$ and $v_0>\pi/2r$. Using similar arguments to those in (A)(1), we have the following. (i) For $0\leqq v\leqq\omega_0$, $\mathrm{Re} f(iv)<0$. (ii) If $\omega_0\leqq v\leqq v_0\leqq\pi/2r$ then $\mathrm{Im} f(iv)>0$. (iii) When $v_0\leqq v\leqq\pi/2r$ it is $\mathrm{Re} f(iv)>0$. (iv) For $\omega_0\leqq v\leqq\pi/2r$ and $v_0>\pi/2r$ we have $\mathrm{Im} f(iv)>0$. (v) When $\pi/2r\leqq v\leqq\pi/r$ then $\mathrm{Re} f(iv)>0$.

Thus when $z=iv$ proceeds along $\Gamma_{\nu 1}$, $f(z)$ leaves the left-half plane and goes to the right one crossing the imaginary axis with $\mathrm{Im} f(z)>0$. For z running along $\Gamma_{\nu 2}^+$, as in (B), we have always $\mathrm{Im} f(z)<0$. When z moves along $\Gamma_{\nu 3}^+$, making $n=0$ in (C), we can see that $f(z)$ returns to the left-half plane crossing the imaginary axis with $\mathrm{Im} f(z)<0$. Finally, as before, for z along $\Gamma_{\nu 0}$, $f(z)$ runs over the negative real axis from $f(c)$ to $f(0)$. Hence $f(\Gamma_\nu^+)$ encircles the origin once, and consequently $f(z)$ has two zeros with positive real part such that $0<|\mathrm{Im} z|<\pi/r$. This achieves the proof of Theorem 3.1.

Notice that when $a=0$, Theorem 3.1 holds for every delay r. In fact, in this case, since we can make $\theta(0)=-\pi/2$, $\theta(iv)$ increases and the alignment ω_0 always exists.

In all other circumstances we are in a situation of instability. As a matter of fact, when $0<k<1$, it can be shown that $f(z)$ has always a real positive zero, and as we have seen, this situation is maintained when $-1<a/b<k<0$. When $-1<k<a/b<0$ and the alignment at some ω_0 in $]0,\pi/2r[$ does not exist, it can be seen that $f(z)$ has two zeros inside $\Gamma_{\pi/r}$.

When $0<\gamma<\alpha$ we have $0<a<b$. It is well known that for $|k|<a/b$, all roots of $f(z)$ lie in the left-half complex plane (see [6]). This can also be proved studying the variation of the function $g(z)$ and applying Rouché's theorem as before. When $0<a/b<k$, $f(z)$ has always a positive real zero, and for $k=a/b$ we have $f(0)=0$. For $-1<k<-a/b$, a statement similar to Theorem 3.1 holds, as was pointed out in [1].

In fact, using the same procedure, it can be shown that there exists always a unique $\omega_1\varepsilon]0,\pi/r[$ such that $f(\pm i\omega_1)=0$ if $|k|=|g(i\omega_1)|$. Therefore if $|k|<|g(i\omega_1)|$, all zeros of $f(z)$ have negative real part. Otherwise $f(z)=0$ for some z having $\mathrm{Re} z>0$ and $0<|\mathrm{Im} z|<\pi/r$.

REFERENCES

[1] R. K. BRAYTON, *Bifurcation of periodic solutions in a nonlinear difference-differential equation of neutral type*, Quart. Appl. Math., 24 (1966), pp. 215–224.

[2] _____, *Nonlinear oscillations in a distributed network*, Quart. Appl. Math., 24 (1967), pp. 289–301.

[3] J. K. HALE, *Theory of Functional Differential Equations*, Applied Mathematical Sciences 3, Springer-Verlag, New York, 1977.

[4] D. HENRY, *Linear autonomous neutral functional differential equations*, J. Differential Equations, 15 (1974), pp. 106–128.

[5] J. M. MAHAFFY, *Periodic solutions for certain protein synthesis models*, J. Math. Anal. Appl., 74 (1980), pp. 72–105.

[6] M. SLEMROD, *Nonexistence of oscillations in a distributed network*, J. Math. Anal. Appl., 36 (1971), pp. 22–40.

SIAM J. MATH. ANAL.
Vol. 17, No. 1, January 1986

006

SEMIGROUPS GENERATED BY A NEUTRAL FUNCTIONAL DIFFERENTIAL EQUATION*

OLOF J. STAFFANS†

Abstract. We discuss a number of semigroups generated by neutral functional differential equations of the form

$$\frac{d}{dt}(x(t)+\mu * x(t))+\nu * x(t)=f(t), \qquad t\geqq 0,$$
$$x(t)=\varphi(t), \qquad t\leqq 0.$$

They are of extended initial function type and of extended forcing function type, and they differ from each other by the amount of smoothness which is imposed on x and f above. The extended initial function type semigroups are adjoints of the extended forcing function type semigroups, and vice versa. The two types of semigroups are also equivalent in the sense that there is a one-to-one, bicontinuous mapping of the state space onto itself, which maps a semigroup of the initial function type onto a semigroup of the forcing function type. In particular, it suffices to study the asymptotic behavior of one of the two types of semigroups, because the results can easily be transferred to the other type of semigroups.

1. Introduction. We discuss a number of semigroups generated by functional differential equations of the form

$$\frac{d}{dt}(x(t)+\mu * x(t))+\nu * x(t)=f(t), \qquad t\geqq 0, \tag{1.1}$$

with initial condition

$$x(t)=\varphi(t), \qquad t\leqq 0. \tag{1.2}$$

The values of x, f and φ lie in $\mathbf{R}^n$, and μ and ν are n by n matrix valued measures on $[0,\infty)$. The measure μ is not allowed to have a point mass at zero. The convolution $\mu * x$ is defined a.e. by

$$\mu * x=\int_{[0,\infty)}[d\mu(s)]\,x(t-s), \qquad t\geqq 0.$$

Our semigroups act on certain "fading memory spaces" of functions of type L^p, or $W^{1,p}$, or $W^{-1,p}$, with $1<p<\infty$, defined on $\mathbf{R}^-=(-\infty,0]$ or $\mathbf{R}^+[0,\infty)$. More specifically, we let η be an "influence function", choose our initial function φ from either $W^{1,p}(\mathbf{R}^-;\mathbf{R}^n;\eta)$ or $L^p(\mathbf{R}^-;\mathbf{R}^n;\eta)$, and choose our forcing function from either $L^p(\mathbf{R}^+;\mathbf{R}^n;\eta)$ or $W^{-1,p}(\mathbf{R}^+;\mathbf{R}^n;\eta)$. The semigroups which we construct are of two types; an extended initial function type, and an extended forcing function type. One gets an extended initial function type semigroup roughly by solving (1.1) with initial condition (1.2), taking a translate of the solution to be a new initial function, and taking a translate of the forcing function f to be the new forcing function. To get an extended forcing function type semigroup we, roughly speaking, solve (1.1) with zero initial condition, and let the new forcing function be a translate of the old forcing function f, plus a correction term which replaces the initial function term in the initial function type semigroup.

We show that the adjoint of a semigroup of extended initial function type is a semigroup of extended forcing function type, and vice versa. We also show that there is

*Received by the editors September 6, 1983, and in revised form March 16, 1984.
†Institute of Mathematics, Helsinki University of Technology, SF-02150 Espoo 15, Finland.

a continuous, one-to-one mapping of the state space onto itself, which maps a initial function type semigroup onto a forcing function type semigroup. In other words, the two types of semigroup are equivalent to each other. This means e.g. that it suffices to study the asymptotic behavior of one of the two types of semigroup, because the other type of semigroups behaves in exactly the same way.

This paper may be considered as a continuation of [33] where the same type of results are proved for a (nondifferentiated) functional equation. We expect the reader to be familiar with [33], and throughout use the same notations as in [33]. We also refer the reader to [33] for a short discussion of earlier comparable results.

In [33, §§5 and 6] we gave two examples on how the equivalence relation between the two types of semigroups could be used to find the generator of the extended forcing function semigroup, and to study the asymptotic behavior of the extended forcing function semigroup. The equivalence relationships which we state here can be used in the same way.

2. The initial function semigroups. In this work we shall for simplicity restrict ourselves to state spaces of type L^p, with $1<p<\infty$. Also, in [33] the limiting cases $p=1$ and $p=\infty$ were discussed. They could be included here, too, but the proofs are slightly simpler in the reflexive case $1<p<\infty$.

We let η be an influence function dominated by a dominating function ρ, and let $L^p(\mathbf{R}^-;\mathbf{R}^n;\eta)$ and $L^p(\mathbf{R}^+;\mathbf{R}^n;\eta)$ be the standard L^p-spaces on $\mathbf{R}^-$ and $\mathbf{R}^+$ with weight η (cf. [33]). Our principal initial function space will be either $L^p(\mathbf{R}^-;\mathbf{R}^n;\eta)$ or $W^{1,p}(\mathbf{R}^-;\mathbf{R}^n;\eta)$, where

$$W^{1,p}(\mathbf{R}^-;\mathbf{R}^n;\eta)=\{\varphi\in L^p(\mathbf{R}^-;\mathbf{R}^n;\eta)|\varphi'\in L^p(\mathbf{R}^-;\mathbf{R}^n;\eta)\}.$$

Here the condition on the derivative φ' should be interpreted as requiring that φ has to be absolutely continuous. Our principal forcing function space will be either $L^p(\mathbf{R}^-;\mathbf{R}^n;\eta)$ or $W^{-1,p}(\mathbf{R}^+;\mathbf{R}^n;\eta)$, a space of distributions which will be defined later. In addition to the principal spaces above, we shall also discuss certain initial and forcing function spaces of cross product type.

The measures μ and ν in (1.1) are throughout supposed to belong to $M(\mathbf{R}^+;\mathbf{R}^{n\times n};\rho)$, and μ is not allowed to have a point mass at zero.

To make a long story short, we define $W^{-1,p}(\mathbf{R}^+;\mathbf{R}^n;\eta)$ to be the dual of the space $W^{1,q}(\mathbf{R}^-;\mathbf{R}^n;\tilde{\eta})$, where q is the conjugate index to p, and $\tilde{\eta}$ is the adjoint influence function to η, i.e. $\tilde{\eta}(t)=[\eta(-t)]^{-1}$, $t\in\mathbf{R}$. We postpone the precise definition of $W^{-1,p}(\mathbf{R}^+;\mathbf{R}^n;\eta)$ to the next section. For the moment it suffices to know that an element of $W^{-1,p}(\mathbf{R}^+;\mathbf{R}^n;\eta)$ can be identified with a distribution of the form $\delta z+f+g'$, where $(z,f,g)\in\mathbf{R}^n\times L^p(\mathbf{R}^+;\mathbf{R}^n;\eta)\times L^p(\mathbf{R}^+;\mathbf{R};\eta)$, δ is the unit point mass at zero, and both f and g vanish on $\mathbf{R}^-$.

Under the general hypothesis, the equation (1.1) has a fundamental solution r, which vanishes on $\mathbf{R}^-$, belongs to $L^1(\mathbf{R}^+;\mathbf{R}^{n\times n};e^{-dt})$ and has a (distribution) derivative in $M(\mathbf{R}^+;\mathbf{R}^{n\times n};e^{-dt})$ for some sufficiently large number d. This function satisfies

$$r'+r'*\mu+r*\nu=r'+\mu*r'+\nu*r=\delta, \tag{2.1}$$

where δ is the unit point mass at zero, and the equation should be inerpreted as an equation in $M(\mathbf{R}^+;\mathbf{R}^{n\times n};e^{-dt})$. See [28, Thm. 5.2]. In particular, $r(0)=I$ (the identity matrix). In general r does not commute with μ and ν. However, by (2.1),

$$(\delta+\mu)*r'*(\delta+\mu)=(\delta+\mu)*(\delta-r*\nu)=(\delta-\nu*r)*(\delta+\mu),$$

so r satisfies

$$(\delta+\mu)*r*\nu=\nu*r*(\delta+\mu). \tag{2.2}$$

Equation (1.1) generates the following initial function type semigroup in $W^{1,p}(\mathbf{R}^-;\mathbf{R}^n;\eta)\times L^p(\mathbf{R}^+;\mathbf{R}^\eta;\eta)$.

THEOREM 2.1. *For each* $(\varphi,f)\in W^{1,p}(\mathbf{R}^-;\mathbf{R}^n;\eta)\times L^p(\mathbf{R}^+;\mathbf{R}^n;\eta)$, *let* $x=x(\varphi,f)$ *be the solution of* (1.1) *with initial condition* (1.2), *and define* $T(t)(\varphi,f)=(x_t,f_t)$, $t\geqq 0$, *where* x_t *is the restriction to* $\mathbf{R}^-$ *of* $\tau_t x$, *and* f_t *is the restriction to* $\mathbf{R}^+$ *of* $\tau_t f$. *Then* T *is a strongly continuous semigroup in* $W^{1,p}(\mathbf{R}^-;\mathbf{R}^n;\eta)\times L^p(\mathbf{R}^+;\mathbf{R}^n;\eta)$.

Here τ_t is the left-translation operator $\tau_t\varphi(s)=\varphi(s+t)$. In the sequel we shall use the subindex t with the same interpretation as in Theorem 1: Whenever a function, call it h, is given the subindex t, then we mean the restriction to either $\mathbf{R}^-$ or to $\mathbf{R}^+$ of $\tau_t h$.

Theorem 2.1 can be deduced from e.g. [30, §7] and [31, §7] (although it is not formulated there in exactly this form). In the situation above one can express the solution x of (1.1) with initial condition (1.2) e.g. in the form

$$x=\varphi+(r+r*\mu)\varphi(0)+r*(f+N(\varphi)+M(\varphi')), \tag{2.3}$$

where we have defined f, $M(\varphi')$ and $N(\varphi)$ to be zero on $(-\infty,0)$, φ and φ' to be zero on $\mathbf{R}^+$, and

$$\begin{aligned} M(\psi)(t)&=-\int_{(t,\infty)} d\mu(s)\psi(t-s), \qquad t\geqq 0,\\ N(\varphi)(t)&=-\int_{(t,\infty)} d\nu(s)\varphi(t-s), \qquad t\geqq 0. \end{aligned} \tag{2.4}$$

To verify that (2.3) is a solution of (1.1) one can simply differentiate (2.3), and use (2.1) and (2.2) above to get (1.1); by [28, §§4 and 7], the solution of (1.1) with initial condition (1.2) is unique under the assumption of Theorem 2.1.

In the sequel we shall throughout use the same convention as in (2.3), i.e. initial functions are always extended to $\mathbf{R}$ by zero on $\mathbf{R}^+$, and forcing functions are always extended to $\mathbf{R}$ by zero on $\mathbf{R}^-$.

Equation (1.1) with initial condition (1.2) generates initial function type semigroups also in other settings. If we relax the smoothness requirement on the initial function to $\varphi\in L^p(\mathbf{R}^-;\mathbf{R}^n;\eta)$, then the forcing function space has to be changed, too, or some other modifications are necessary. One possible way, the one used in [28], is to rewrite (1.1) in the form

$$\begin{aligned} z'(t)+\nu*x(t)&=f(t), \qquad t\geqq 0,\\ x(t)+\mu*x(t)-z(t)&=g(t), \qquad t\geqq 0, \end{aligned} \tag{2.5}$$

and replace the initial condition by

$$x(t)=\varphi(t), \quad t<0, \qquad z(0)=\zeta. \tag{2.6}$$

Then one gets the following semigroup.

THEOREM 2.2. *For each* $(\varphi,\zeta,f,g)\in L^p(\mathbf{R}^-;\mathbf{R}^n;\eta)\times\mathbf{R}^n\times L^p(\mathbf{R}^+;\mathbf{R}^n;\eta)\times L^p(\mathbf{R}^+;\mathbf{R}^n;\eta)$, *let* $(x,z)=(x(\varphi,\zeta,f,g),\ z(\varphi,\zeta,f,g))$ *be the solution of* (2.5) *with initial condition* (2.6), *and define* $T(t)(\varphi,\zeta,f,g)=(x_t,z(t),f_t,g_t)$, $t\geqq 0$. *Then* T *is a strongly continuous semigroup in* $L^p(\mathbf{R}^-;\mathbf{R}^n;\eta)\times\mathbf{R}^n\times L^p(\mathbf{R}^+;\mathbf{R}^n;\eta)\times L^p(\mathbf{R}^+;\mathbf{R}^n;\eta)$.

Here, the solution (x,z) of (2.5) with initial condition (2.6) is given by

$$\begin{aligned} x &= \varphi + r\zeta + r*(f+N(\varphi)) + r'*(g+M(\varphi)), \\ z &= (r+\mu*r)\zeta + (r+\mu*r)*(f+N(\varphi)) - \nu*r*(g+M(\varphi)), \end{aligned} \tag{2.7}$$

where $M(\varphi)$ and $N(\varphi)$ are defined as in (2.4) (and initial and forcing functions are defined to be zero outside of their original domain).

Although it is true that x and z together determine the functions f and g in (2.5) uniquely, it is not true that x alone does so. This is a rather disturbing fact, because we only added the new variable z in order to make the problem "well posed", and it is not necessarily true that z has a natural physical interpretation. In other words there is a certain redundancy in the forcing function pair (f,g) in (2.5). One can remove this redundancy in many ways. One way is very obvious: If $g\equiv 0$, then the g-component of the semigroup remains zero, and by restricting the previous semigroup to the space $L^p(\mathbf{R}^-;\mathbf{R}^n;\eta)\times\mathbf{R}^n\times L^p(\mathbf{R}^+;\mathbf{R}^n;\eta)\times\{0\}$ we get the following semigroup.

THEOREM 2.3. *For each* $(\varphi,\zeta,f)\in L^p(\mathbf{R}^-;\mathbf{R}^n;\eta)\times\mathbf{R}^n\times L^p(\mathbf{R}^+;\mathbf{R}^n;\eta)$, *let* $(x,z)=(x(\varphi,\zeta,f),z(\varphi,\zeta,f))$ *be the solution of* (2.5) *with* $g\equiv 0$, *with initial condition* (2.6). *Define* $T(t)(\varphi,\zeta,f)=(x_t,z(t),f_t)$. *Then* T *is a strongly continuous semigroup in* $L^p(\mathbf{R}^-;\mathbf{R}^n;\eta)\times\mathbf{R}^n\times L^p(\mathbf{R}^+;\mathbf{R}^n;\eta)$.

Another equally obvious possibility is to take the f-component of the semigroup to be identically zero. This is roughly equivalent to replacing $f(t)$ in (1.1) by $df(t)/dt$.

Theorem 2.2 has the obvious disadvantage that the set of permitted forcing functions is much more restricted than the set of forcing functions in Theorem 2.2. Fortunately there is also another way of removing the redundancy in Theorem 2.2, which leaves the set of forcing functions intact. Interpreting (2.7) in the distribution sense, we can write x in the form

$$x=\varphi+r*N(\varphi)+r'*M(\varphi)+r*f, \tag{2.8}$$

where we have replaced $\delta z+f+g'$ in (2.7) by a distribution $f\in W^{-1,p}(\mathbf{R}^+;\mathbf{R}^n;\eta)$. In Theorem 2.2 we have used a particular representation for $f\in W^{-1,p}(\mathbf{R}^+;\mathbf{R}^n;\eta)$, but the solution x depends only on the distribution f, and not on the particular representation. Therefore, we can interpret Theorem 2.2 in the following way.

THEOREM 2.4. *For each* $(\varphi,f)\in L^p(\mathbf{R}^-;\mathbf{R}^n;\eta)\times W^{-1,p}(\mathbf{R}^+;\mathbf{R}^n;\eta)$, *choose an arbitrary triple* $(\zeta,h,g)\in\mathbf{R}^n\times L^p(\mathbf{R}^+;\mathbf{R}^n;\eta)\times L^p(\mathbf{R}^+;\mathbf{R}^n;\eta)$ *representing* f, *in the sense that* $f=\delta\zeta+h+g'$, *and let* $(x,z)=(x(\varphi,\zeta,h,g),z(\varphi,\zeta,h,g))$ *be the solution of* (2.5) *with* f *replaced by* h, *and with initial condition* (2.6). *Define* $T(t)(\varphi,f)=(x_t,f_t)$, $t\geqq 0$, *where* $f_t=\delta z(t)+h_t+(g_t)'$. *Then* $T(t)$ *is a strongly continuous semigroup in* $L^p(\mathbf{R}^-;\mathbf{R}^n;\eta)\times W^{-1,p}(\mathbf{R}^+;\mathbf{R}^n;\eta)$.

3. The adjoints of the rough initial function semigroups. We next want to compute the adjoints of the initial function semigroups found above. We begin with the adjoints of the "rough" semigroups in Theorems 2.2–2.4 (the adjoint of the semigroup in Theorem 2.1 will be discussed in §4).

In all the different cases considered above, the state space of the initial function type semigroups are reflexive. Therefore, the adjoint semigroups can simply be computed as the adjoints of the original semigroups (it is not necessary to restrict the dual space to the closure of the domain of the adjoint of the generator).

Let us first compute the adjoint of the semigroup in Theorem 2.2. The dual space of the state space can in this case be identified with $L^q(\mathbf{R}^-;\mathbf{R}^n;\tilde{\eta})\times L^q(\mathbf{R}^-;\mathbf{R}^n;\tilde{\eta})\times\mathbf{R}^n \times L^q(\mathbf{R}^+;\mathbf{R}^n;\tilde{\eta})$ through the duality mapping

$$\big\langle (f^*,g^*,\zeta^*,\varphi^*),(\varphi,\zeta,f,g)\big\rangle = \varphi^**\varphi(0)+\zeta^*\zeta+f^**f(0)+g^**g(0).$$

Here q is the conjugate index to p, and $\tilde{\eta}$ is the adjoint influence function to η, i.e. $\tilde{\eta}(t)=[\eta(-t)]^{-1}$, $t\in\mathbf{R}$. If we denote $T^*(f^*,g^*,\zeta^*,\varphi^*)(t)$ by $(\tilde{f},\tilde{g},\tilde{\zeta},\tilde{\varphi})$, then a straightforward computation, very similar to the corresponding one in [33], leads to the following equations for $\tilde{f},\tilde{g},\tilde{\zeta}$, and $\tilde{\varphi}$ (in particular, the adjoints of the operators M and N are computed in the same way as the adjoint of the operator G in [33]):

$$\begin{aligned}
\tilde{f} &= \tau_t\big[f^*+\zeta^*(r+\mu*r)+\varphi^**r\big],\\
\tilde{g} &= \tau_t\big[g^*-\zeta^*(\nu*r)+\varphi^**r'\big],\\
\tilde{\zeta} &= \zeta^*(r+\mu*r)(t)+\varphi^**r(t),\\
\tilde{\varphi} &= \tau_t\big[\varphi^*+N^*\big(\zeta^*(r+\mu*r)+\varphi^**r\big)\\
&\qquad +M^*\big(-\zeta^*(\nu*r)+\varphi^**r'\big)\big],
\end{aligned}$$

where M^* and N^* are defined analogously to M and N, namely

$$M^*g^*(s) = -\int_{(t,\infty)} g^*(t-s)\,d\mu(s), \qquad t\geqq 0,$$

$$N^*f^*(t) = -\int_{(t,\infty)} f^*(t-s)\,d\nu(s), \qquad t\geqq 0.$$

These equations can be interpreted in the following way. Define (x^*,z^*) by

$$\begin{aligned}
x^* &= f^*+\zeta^*(r+\mu*r)+\varphi^**r,\\
z^* &= g^*-\zeta^*(\nu*r)+\varphi^*+r'.
\end{aligned} \tag{3.1}$$

Then $x^*\in L^q(\mathbf{R}^-;\mathbf{R}^n;\tilde{\eta})\cap W^{1,q}(\mathbf{R}^+;\mathbf{R}^n;\tilde{\eta})$, $z^*\in L^q(\mathbf{R},\mathbf{R}^n,\tilde{\eta})$, and $(x^*)'(t)=-\zeta^*(\nu*r)(t)+\varphi^**r'(t)=z^*(t)$ for almost all $t\geqq 0$. This means that (x^*,z^*) is the solution of the equations

$$\begin{aligned}
&(x^*)'(t)+\int_{(0,t]}(x^*)'(t-s)\,d\mu(s)+\int_{[0,t]}x^*(t-s)\,d\nu(s)=\varphi^*(t), && t\geqq 0,\\
&z^*(t)=(x^*)'(t), && t\geqq 0,
\end{aligned} \tag{3.2}$$

with initial conditions

$$x^*(t)=f^*(t),\quad z^*(t)=g^*(t),\quad t<0, \qquad x^*(0)=\zeta^*. \tag{3.3}$$

Moreover, $\tilde{f}=x_t^*$, $\tilde{g}=z_t^*$, $\tilde{\zeta}=x^*(t)$, and $\tilde{\varphi}=\varphi_t^*+N^*(x_t^*-f_t^*)+M^*(z_t^*+g_t^*)$. In other words, the adjoint semigroup can be described as follows.

THEOREM 3.1. *The state space of the adjoint semigroup T^* of the semigroup T in Theorem* 2.2 *is $L^q(\mathbf{R}^-,\mathbf{R}^n,\tilde{\eta})\times L^q(\mathbf{R}^-,\mathbf{R}^n,\tilde{\eta})\times\mathbf{R}^n\times L^q(\mathbf{R}^+;\mathbf{R}^n;\tilde{\eta})$, and T^* is given by*

$$\begin{aligned}
&T^*(f^*,g^*,\zeta^*,\varphi^*)(t)\\
&\quad=\big(x_t^*,z_t^*,x^*(t),\varphi_t^*+N^*\big(x_t^*-f_t^*\big)+M^*\big(z_t^*-g_t^*\big)\big),
\end{aligned}$$

where $(x^*,z^*)=\big(x^*(f^*,g^*,\zeta^*,\varphi^*),z^*(f^*,g^*,\zeta^*,\varphi^*)\big)$ *is the solution of* (3.2) *with initial condition* (3.3), *and*

$$
\begin{aligned}
M^*\big(z_t^*-g_t^*\big)(s)&=-\int_{(s,s+t]}(x^*)'(t+s-v)\,d\mu(v), \qquad s\geqq 0,\\
N^*\big(x_t^*+f_t^*\big)(s)&=-\int_{(s,s+t]}x^*(t+s-v)\,d\nu(v), \qquad s\geqq 0.
\end{aligned}
\tag{3.4}
$$

The adjoint of the semigroup in Theorem 2.3 is obtained from the semigroup above, by simply deleting the g^*-component (it is possible to delete the g^*-component, because all the other components are independent of g^*; in particular, by (3.4), $M^*(z_t^*-g_t^*)$ depends only on $(x^*)'$, and not on g^*).

THEOREM 3.2. *The state space of the adjoint semigroup of the semigroup in Theorem* 2.3 *is* $L^q(\mathbf{R}^-;\mathbf{R}^n;\tilde{\eta})\times\mathbf{R}^n\times L^q(\mathbf{R}^+;\mathbf{R}^n;\tilde{\eta})$, *and it is obtained from the semigroup in Theorem* 3.1 *through a deletion of the* g^**-component.*

Before applying Theorem 3.1 to get the adjoint of the semigroup in Theorem 2.4, let us give a precise definition of $W^{-1,p}(\mathbf{R}^+;\mathbf{R}^n;\eta)$. As we already mentioned above, we define this space to be the dual space of $W^{1,q}(\mathbf{R}^-;\mathbf{R}^n;\tilde{\eta})$. More specifically, we imbed $W^{1,p}(\mathbf{R}^-;\mathbf{R}^n;\eta)$ into $L^p(\mathbf{R}^-;\mathbf{R}^n;\eta)\times L^p(\mathbf{R}^-;\mathbf{R}^n;\eta)\times\mathbf{R}^n$, identifying $W^{1,p}(\mathbf{R}^-;\mathbf{R}^n;\eta)$ with the subspace

$$\big\{\big(\varphi,\varphi',\varphi(0)\big)\big|\varphi\in W^{1,p}(\mathbf{R}^-;\mathbf{R}^n;\eta)\big\}$$

of $L^p(\mathbf{R}^-;\mathbf{R}^n;\eta)\times L^p(\mathbf{R}^-;\mathbf{R}^n;\eta)\times\mathbf{R}^n$. In the same way as imbed $W^{1,q}(\mathbf{R}^-;\mathbf{R}^n;\tilde{\eta})$ into $L^q(\mathbf{R}^-;\mathbf{R}^n;\tilde{\eta})\times L^q(\mathbf{R}^-;\mathbf{R}^n;\tilde{\eta})\times\mathbf{R}^n$. Every continuous linear functional on $L^q(\mathbf{R}^-;\mathbf{R}^n;\tilde{\eta})\times L^q(\mathbf{R}^-;\mathbf{R}^n;\tilde{\eta})\times\mathbf{R}^n$ can be represented by a triple $(z,f,g)\in\mathbf{R}^n\times L^p(\mathbf{R}^+;\mathbf{R}^n;\eta)\times L^p(\mathbf{R}^+;\mathbf{R}^n;\eta)$ through the formula

$$\big\langle(z,f,g),(\varphi,\psi,\xi)\big\rangle=z\xi+f*\varphi(0)+g*\psi(0).$$

In particular, every continuous linear functional φ^* on $W^{1,q}(\mathbf{R}^-;\mathbf{R}^n;\tilde{\eta})$ can also be represented by a triple $(z,f,g)\in\mathbf{R}^n\times L^p(\mathbf{R}^+;\mathbf{R}^n;\eta)\times L^p(\mathbf{R}^+;\mathbf{R}^n;\eta)$ through the formula

$$\big\langle\varphi^*,\varphi\big\rangle=z\varphi(0)+f*\varphi(0)+g*\varphi'(0). \tag{3.5}$$

Two functionals φ_1^* and φ_2^* induced by z_1,f_1,g_1 and z_2,f_2,g_2, respectively, are identical if and only if $z_2=z_1-h(0)$, $f_2=f_1-h'$ and $g_2=g_1+h$ for some function $h\in W^{1,p}(\mathbf{R}^+;\mathbf{R}^n;\eta)$. In other words, the dual space of $W^{1,q}(\mathbf{R}^-;\mathbf{R}^n;\tilde{\eta})$ can be regarded as the quotient space of $\mathbf{R}^n\times L^p(\mathbf{R}^+;\mathbf{R}^n;\eta)\times L^p(\mathbf{R}^+;\mathbf{R}^n;\eta)$ over its subspace

$$\big\{\big(-h(0),-h',h\big)\big|h\in W^{1,p}(\mathbf{R}^+;\mathbf{R}^n;\eta)\big\}.$$

(Of course, this subspace can also be identified with $W^{1,p}(\mathbf{R}^+;\mathbf{R}^n;\eta)$, but we shall not use that identification here; instead we interpret it to be the orthogonal complement to $W^{1,q}(\mathbf{R}^-,\mathbf{R}^n;\tilde{\eta})$ in $\mathbf{R}^n\times L^p(\mathbf{R}^+;\mathbf{R}^n;\eta)\times L^p(\mathbf{R}^+;\mathbf{R}^n;\eta)$.) As we mentioned above, we shall denote this quotient space by $W^{-1,p}(\mathbf{R}^+;\mathbf{R}^n;\eta)$.

(The preceding notation is not completely standard. Frequently the notation $W^{-1,p}(\mathbf{R}^+;\mathbf{R}^n;\eta)$ is used for the dual space of $W_0^{1,q}(\mathbf{R}^-;\mathbf{R}^n;\tilde{\eta})$, where $W_0^{1,q}(\mathbf{R}^-;\mathbf{R}^n;\tilde{\eta})=\{\varphi\in W^{1,q}(\mathbf{R}^-;\mathbf{R}^n;\tilde{\eta})\,|\,\varphi(0)=0\}$. One can get the latter space from the former by taking the quotient of $W^{-1,p}(\mathbf{R}^+;\mathbf{R}^n;\eta)$ over the subspace of functionals in $W^{-1,p}(\mathbf{R}^+;\mathbf{R}^n;\eta)$ which are of the form (3.5) with $f\equiv g\equiv 0$.)

There is a simple distribution interpretation of (3.5) which we already used in the formulation of Theorem 2.4. Define $f(t)=g(t)=0$, $t<0$, and extend φ to a function in $W^{1,p}(\mathbf{R};\mathbf{R}^n;\eta)$ in an arbitrary way. Then $\langle\varphi^*,\varphi\rangle=(\delta z+f+g')*\varphi(0)$, where δ is the unit point mass at zero. This means that in the distribution sense, $\varphi^*=\delta z+f+g'$.

Let us go back to the adjoint of the semigroup in Theorem 2.4. To get this adjoint we have to restrict the semigroup in Theorem 3.1 to the dual space of $L^p(\mathbf{R}^-;\mathbf{R}^n;\eta)\times W^{-1,p}(\mathbf{R}^+;\mathbf{R}^n;\eta)$, which is $W^{1,q}(\mathbf{R}^-;\mathbf{R}^n;\tilde{\eta})\times L^q(\mathbf{R}^-;\mathbf{R}^n;\tilde{\eta})$. The latter space we have identified with the product of the subspace

$$\{(f^*,(f^*)',f^*(0))|f^*\in W^{1,q}(\mathbf{R}^-;\mathbf{R}^n;\tilde{\eta})\}$$

of $L^q(\mathbf{R}^-;\mathbf{R}^n;\tilde{\eta})\times L^q(\mathbf{R}^-;\mathbf{R}^n;\tilde{\eta})\times\mathbf{R}^n$ and $L^q(\mathbf{R}^-;\mathbf{R}^n;\tilde{\eta})$. If we drop the g^*- and ζ^*-components of the semigroup in Theorem 3.1, replacing g^* by $(f^*)'$ and ζ^* by $f^*(0)$, then (3.2) and (3.3) become

$$(3.6)\quad (x^*)'(t)+\int_{(0,t]}(x^*)'(t-s)\,d\mu(s)+\int_{[0,t]}x^*(t-s)\,d\nu(s)=\varphi^*(t),\qquad t\geqq 0,$$

$$(3.7)\quad x^*(t)=f^*(t),\qquad t\leqq 0,$$

and we get the following semigroup in $W^{1,q}(\mathbf{R}^-,\mathbf{R}^n;\tilde{\eta})\times L^q(\mathbf{R}^-;\mathbf{R}^n;\tilde{\eta})$.

THEOREM 3.3. *The state space of the adjoint semigroup T^* of the semigroup T in Theorem* 2.4 *is $W^{1,q}(\mathbf{R}^-;\mathbf{R}^n;\tilde{\eta})\times L^q(\mathbf{R}^+;\mathbf{R}^n;\tilde{\eta})$, and T^* is given by*

$$T^*(f^*,\varphi^*)(t)=\left(x_t^*,\varphi_t^*+N^*\left(x_t^*-f_t^*\right)+M^*\left(\left(x_t^*\right)'-\left(f_t^*\right)'\right)\right),$$

where $x^*=x^*(f^*,\varphi^*)$ is the solution of (3.6) *with initial condition* (3.7), *and the terms containing M^* and N^* are defined as in* (3.4).

We can rewrite (3.1)–(3.4) and (3.6), using the same type of notation as was used in §2, to get the following set of equations:

$$(3.8)\qquad \begin{aligned} x&=\varphi+(r+r*\mu)\xi+r*f,\\ y&=\psi-(r*\nu)\xi+r'*f,\end{aligned}$$

$$(3.9)\qquad \begin{aligned} &x'(t)+\int_{(0,t]}[d\mu(s)]x'(t-s)+\int_{[0,t]}[d\nu(s)]x(t-s)=f(t),\qquad t\geqq 0,\\ &y(t)=x'(t),\qquad t\geqq 0,\end{aligned}$$

$$(3.10)\qquad x(t)=\varphi(t),\ y(t)=\psi(t),\quad t<0,\qquad x(0)=\xi,$$

$$(3.11)\qquad \begin{aligned} M(y_t-\psi_t)&=-\int_{(s,s+t]}[d\mu(v)]x'(t+s-v),\qquad s\geqq 0,\\ N(x_t-\varphi_t)&=-\int_{(s,s+t]}[d\nu(v)]x(t+s-v),\qquad s\geqq 0,\end{aligned}$$

$$(3.12)\quad x'(t)+\int_{(0,t]}[d\mu(s)]x'(t-s)+\int_{[0,t]}[d\nu(x)](t-s)=f(t),\qquad t\geqq 0.$$

With the new notation the semigroup which we found in Theorems 3.1–3.3 can be described as follows.

THEOREM 3.4. *For each* $(\varphi,\psi,\xi,f)\in L^p(\mathbf{R}^-;\mathbf{R}^n;\eta)\times L^p(\mathbf{R}^-;\mathbf{R}^n;\eta)\times\mathbf{R}^n\times L^p(\mathbf{R}^+;\mathbf{R}^n;\eta)$, *let* $(x,y)=(x(\varphi,\psi,\xi,f),y(\varphi,\psi,\xi,f))$ *be the solution of* (3.9) *with initial condition* (3.10), *and define*

$$S(t)(\varphi,\psi,\xi,f)=(x_t,y_t,x(t),f_t+N(x_t-\varphi_t)+M(y_t+\psi_t)),$$

where $M(y_t-\psi_t)$ *and* $N(x_t-\varphi_t)$ *are defined as in* (3.11). *Then* S *is a strongly continuous semigroup in* $L^p(\mathbf{R}^-;\mathbf{R}^n;\eta)\times L^p(\mathbf{R}^-;\mathbf{R}^n;\eta)\times\mathbf{R}^n\times L^p(\mathbf{R}^+;\mathbf{R}^n;\eta)$.

THEOREM 3.5. *For each* $(\varphi,\xi,f)\in L^p(\mathbf{R}^-;\mathbf{R}^n;\eta)\times\mathbf{R}^n\times L^p(\mathbf{R}^+;\mathbf{R}^n;\eta)$, *define* $S(t)(\varphi,\xi,f)$ *as above, ignoring the* ψ*-component. Then* S *is a strongly continuous semigroup in* $L^p(\mathbf{R}^-;\mathbf{R}^n;\eta)\times\mathbf{R}^n\times L^p(\mathbf{R}^+;\mathbf{R}^n;\eta)$.

THEOREM 3.6. *For each* $(\varphi,f)\in W^{1,p}(\mathbf{R}^-;\mathbf{R}^n;\eta)\times L^p(\mathbf{R}^+;\mathbf{R}^n;\eta)$, *let* $x=x(\varphi,f)$ *be the solution of* (3.12) *with initial condition* (1.2), *and define*

$$S(t)(\varphi,f)=(x_t,f_t+N(x_t-\varphi_t)+M((x_t)'-(f_t)')),$$

where $M((x_t)'-(f_t)')$ *and* $N(x_t-\varphi_t)$ *are defined as in* (3.11). *Then* S *is a strongly continuous semigroup in* $W^{1,p}(\mathbf{R}^-;\mathbf{R}^n;\eta)\times L^p(\mathbf{R}^+;\mathbf{R}^n;\eta)$.

4. The adjoint of the smooth initial function semigroup. We have still not computed the adjoint of the "smooth" initial function semigroup in Theorem 2.1. The dual space to the state space $W^{1,p}(\mathbf{R}^-;\mathbf{R}^n;\eta)\times L^p(\mathbf{R}^+;\mathbf{R}^n;\eta)$ is $L^q(\mathbf{R}^-;\mathbf{R}^n;\tilde\eta)\times W^{-1,q}(\mathbf{R}^+;\mathbf{R}^n;\tilde\eta)$, so we again have to work in a Sobolev space with negative index. Thinking of $W^{-1,q}(\mathbf{R}^+;\mathbf{R}^n;\tilde\eta)$ as a quotient of $\mathbf{R}^n\times L^q(\mathbf{R}^+;\mathbf{R}^n;\tilde\eta)\times L^q(\mathbf{R}^+;\mathbf{R}^n;\tilde\eta)$, whose dual space is $L^p(\mathbf{R}^-;\mathbf{R}^n;\eta)\times L^p(\mathbf{R}^-;R^n;\eta)\times\mathbf{R}^n$, and looking at the semigroup presented in Theorem 3.6, one soon discovers that the initial function semigroup in Theorem 2.1 can be imbedded in a larger initial function semigroup, which acts on $L^p(\mathbf{R}^-;\mathbf{R}^n;\eta)\times L^p(\mathbf{R}^-;\mathbf{R}^n;\eta)\times\mathbf{R}^n\times L^p(\mathbf{R}^+;\mathbf{R}^n;\eta)$. Basically, to get this semigroup one prescribes two initial functions in (1.1), i.e. one uses the initial function φ in the term $\nu * x$, but replaces φ' by a new initial function ψ in the term $\mu * x'$. The precise formulation resembles the formulation (3.9)–(3.10). We require (x,y) to satisfy

$$\begin{aligned} &y(t)+\mu * y(t)+\nu * x(t)=f(t), \qquad t\geqq 0,\\ &x'(t)=y(t), \qquad t\geqq 0, \end{aligned} \tag{4.1}$$

and use the initial condition (3.10). Here $(\varphi,\psi,\xi,f)\in L^p(\mathbf{R}^-;\mathbf{R}^n;\eta)\times L^p(\mathbf{R}^-;\mathbf{R}^n;\eta)\times\mathbf{R}^n\times L^p(\mathbf{R}^+;\mathbf{R}^n;\eta)$. Using the fundamental solution r in (2.1), one finds that this set of equations has the unique solution

$$\begin{aligned} x&=\varphi+(r+r*\mu)\xi+r*(f+N(\varphi)+M(\psi)),\\ y&=\psi-(r*\nu)\xi+r'*(f+N(\varphi)+M(\psi)) \end{aligned} \tag{4.2}$$

(cf. (2.3) and (3.8)). Translating all functions to the left we get the following semigroup.

THEOREM 4.1. *For each* $(\varphi,\psi,\xi,f)\in L^p(\mathbf{R}^-;\mathbf{R}^n;\eta)\times L^p(\mathbf{R}^-;\mathbf{R}^n;\eta)\times\mathbf{R}^n\times L^p(\mathbf{R}^+;\mathbf{R}^n;\eta)$, *let* $(x,y)=(x(\varphi,\psi,\xi,f),y(\varphi,\psi,\xi,f))$ *be the solution of* (4.1) *with initial condition* (3.10). *Then* T *is a strongly continuous semigroup in* $L^p(\mathbf{R}^-;\mathbf{R}^n;\eta)\times L^p(\mathbf{R}^-;\mathbf{R}^n;\eta)\times\mathbf{R}^n\times L^p(\mathbf{R}^+;\mathbf{R}^n;\eta)$.

By taking $x\in W^{1,p}(\mathbf{R}^-;\mathbf{R}^n;\eta)$, $y=x'$, $\xi=x(0)$, we get the semigroup in Theorem 2.1.

Now, just as in the preceding section, it is a straightforward task to compute the adjoint of the operator $T(t)$ in Theorem 4.1. If we denote $T^*(f^*,\xi^*,\varphi^*,\psi^*)(t)$ by $(\tilde{f},\tilde{\xi},\tilde{\varphi},\tilde{\psi})$, then one gets the following equations:

$$\begin{aligned}
\tilde{f}&=\tau_t[f^*+\xi^*r+\varphi^**r+\psi^**r')],\\
\tilde{\xi}&=\xi^*(r+r*\mu)+\varphi^**r+\varphi^**r*\mu-\psi^**r*\nu,\\
\tilde{\varphi}&=\tau_t[\varphi^*+N^*(\xi^*r+\varphi^**r+\psi^**r')],\\
\tilde{\psi}&=\tau_t[\psi^*+M^*(\xi^*r+\varphi^**r+\psi^**r')].
\end{aligned}$$

One can interpret these equations in the following way. Define

$$\begin{aligned}
x^*&=f^*+\xi^*r+\varphi^**r+\psi^**r',\\
y^*&=\xi^*(r+r*\mu)+\varphi^**(r+r*\mu)-\psi^**r*\nu.
\end{aligned}\tag{4.3}$$

Then $x^*\in L^p(\mathbf{R}^-;\mathbf{R}^n;\eta)$, $y^*\in W^{1,p}(\mathbf{R}^+;\mathbf{R}^n;\eta)$, and (x^*,y^*) is the solution of the equation

$$\begin{aligned}
&(y^*)'(t)+\int_{[0,t]}x^*(t-s)\,d\nu(s)=\varphi^*(t),\qquad t\geqq 0,\\
&x^*(t)+\int_{(0,t]}x^*(t-s)\,d\mu(s)-y^*(t)=\psi^*(t),\qquad t\geqq 0,
\end{aligned}\tag{4.4}$$

with initial condition

$$x^*(t)=f^*,\quad t<0,\qquad y^*(0)=\xi^*.\tag{4.5}$$

Moreover, $\tilde{f}=x_t^*$, $\tilde{\xi}=y^*(t)$, $\tilde{\varphi}=\varphi_t^*-N^*(x_t^*-f_t^*)$, and $\tilde{\psi}=\psi_t^*+M^*(x_t^*-f_t^*)$. In other words, the adjoint of the semigroup in Theorem 4.1 can be described as follows.

THEOREM 4.2. *The state space of the adjoint semigroup T^* of the semigroup T in Theorem* 4.1 *is $L^q(\mathbf{R}^-;\mathbf{R}^n;\bar{\eta})\times\mathbf{R}^n\times L^q(\mathbf{R}^+;\mathbf{R}^n;\tilde{\eta})\times L^q(\mathbf{R}^+;\mathbf{R}^n;\tilde{\eta})$, and T^* is given by*

$$T^*(f^*,\xi^*,\varphi^*,\psi^*)=\big(x_t^*,y^*(t),\varphi_t^*+N^*(x^*{}_t-f^*{}_t),\psi^*{}_t+M^*\big(x_t^*-f_t^*\big)\big),$$

where $(x^*,y^*)=(x^*(f^*,\xi^*,\varphi^*,\psi^*),y^*(f^*,\xi^*,\varphi^*,\psi^*))$ is the solution of (4.4) *with initial condition* (4.5), *and the terms involving M^* and N^* are defined as in* (3.4) *(with $(x^*)'$ replaced by x^*).*

The definition of f^* in (4.3) can be interpreted in the distribution sense to mean

$$x^*=f^*+\varphi^**r,\tag{4.6}$$

where φ^* stands for the distribution $\xi^*\delta+\varphi^*+(\psi^*)'$ in (4.3). This distribution is uniquely determined as an element of the dual space $W^{-1,q}(\mathbf{R}^+;\mathbf{R}^n;\tilde{\eta})$ of $W^{1,p}(\mathbf{R}^+;\mathbf{R}^n;\eta)$. In the same way we can interpret the definition of $T^*(t)$ in the distribution sense, and we find that the adjoint of the semigroup in Theorem 2.1 can be described as follows.

THEOREM 4.3. *The state space of the adjoint semigroup T^* of the semigroup T in Theorem* 2.1 *is $L^q(\mathbf{R}^-;\mathbf{R}^n;\tilde{\eta})\times W^{-1,q}(\mathbf{R}^+,\mathbf{R}^n;\tilde{\eta})$, and it is defined as follows. For each $(f^*,\varphi^*)\in L^q(\mathbf{R}^-;\mathbf{R}^n;\tilde{\eta})\times W^{-1,q}(\mathbf{R}^+;\mathbf{R}^n;\tilde{\eta})$, choose an arbitrary triple (ξ^*,γ^*,ψ^*) representing φ^* in the sense that $\varphi^*=\xi^*\delta+\gamma^*+(\psi^*)'$, let $(x^*,y^*)=(x^*(f^*,\xi^*,\gamma^*,\psi^*),y^*(f^*,\xi^*,\gamma^*,\psi^*))$ be the solution of* (4.4) *with φ^* replaced by γ^*, and with initial condition* (4.5). *Define $T^*(f^*,\varphi^*)=(x_t^*,\varphi_t^*)$, where $\varphi_t^*\in W^{-1,q}(\mathbf{R}^+;\mathbf{R}^n;\tilde{\eta})$ is the distribution $y^*(t)\delta+\gamma_t^*+N^*(x_t^*-f_t^*)+[\psi_t^*+M^*(x_t^*-f_t^*)]'$.*

Just like in §3, let us again change the notation and describe the last two semigroups in terms of equations directly related to (1.1). Equations (4.3), (4.4) and (4.6) become

$$\begin{aligned} x &= \varphi + r\zeta + r * f + r' * g, \\ z &= (r + \mu * r)\zeta + (r + \mu * r) * f - \nu * r * g, \end{aligned} \tag{4.7}$$

$$\begin{aligned} &z'(t) + \int_{[0,t]} [d\nu(s)]\, x(t-s) = f(t), \qquad t \geqq 0, \\ &x(t) + \int_{(0,t]} [d\mu(s)]\, x(t-s) - z(t) = g(t), \qquad t \geqq 0, \end{aligned} \tag{4.8}$$

$$x = \varphi + r * f, \tag{4.9}$$

and (4.5) becomes (2.6). Observe that apart from the initial function corrections to the forcing functions, formulas (4.7) and (4.9) agree with formulas (2.7) and (2.8).

With the new notation, the semigroups in Theorems 4.2 and 4.3 can be described as follows.

THEOREM 4.4. *For each* $(\varphi,\zeta,f,g) \in L^p(\mathbf{R}^-;\mathbf{R}^n;\eta) \times \mathbf{R}^n \times L^p(\mathbf{R}^+;\mathbf{R}^n;\eta) \times L^p(\mathbf{R}^+;\mathbf{R}^n;\eta)$, *let* $(x,z) = (x(\varphi,\zeta,f,g),\ z(\varphi,\zeta,f,g))$ *be the solution of* (4.8) *with initial condition* (2.6), *and define* $S(t)(\varphi,\zeta,f,g) = (x_t, z(t), f_t + N(x_t - \varphi_t),\ g_t + M(x_t - \varphi_t))$, $t \geqq 0$. *Then* S *is a strongly continuous semigroup in* $L^p(\mathbf{R}^-;\mathbf{R}^n;\eta) \times \mathbf{R}^n \times L^p(\mathbf{R}^+;\mathbf{R}^n;\eta) \times L^p(\mathbf{R}^+;\mathbf{R}^n;\eta)$.

THEOREM 4.5. *For each* $(\varphi,f) \in L^p(\mathbf{R}^-;\mathbf{R}^n;\eta) \times W^{-1,p}(\mathbf{R}^+;\mathbf{R}^n;\eta)$, *choose an arbitrary triple* $(\zeta,h,g) \in \mathbf{R}^n \times L^p(\mathbf{R}^+;\mathbf{R}^n;\eta) \times L^p(\mathbf{R}^+;\mathbf{R}^n;\eta)$ *representing* f *in the sense that* $f = \delta\zeta + h + g'$, *and let* $(x,z) = (x(\varphi,\zeta,h,g), z(\varphi,\zeta,h,g))$ *be the solution of* (4.8) *with* f *replaced by* h, *and with initial condition* (2.6). *Define* $S(t)(\varphi,f) = (x_t, f_t)$, $t \geqq 0$, *where* $f_t = \delta z(t) + h_t + N(x_t - \varphi_t) + [g_t + M(x_t - \varphi_t)]'$. *Then* S *is a strongly continuous semigroup in* $L^p(\mathbf{R}^-;\mathbf{R}^n;\eta) \times W^{-1,p}(\mathbf{R}^+;\mathbf{R}^n;\eta)$.

5. Initial and forcing function semigroups are equivalent. By now we have described a total of ten semigroups generated by (1.1) (plus five more adjoint semigroups). Five of them are of initial function type (the initial function does affect the values of the solution on $\mathbf{R}^+$), and five of them of forcing function type (the initial function does not affect the values of the solution on $\mathbf{R}^+$). Fortunately, one need not study all of them separately, because some of them are similar to each other. In particular, four of the initial function type semigroups are similar to four forcing function type semigroups. In the retarded case also the fifth pair of semigroups is similar. In all cases the similarity transformation is essentially the same, i.e. the operator which adds an initial function correction to the forcing function, as required by (1.1) in its different versions.

THEOREM 5.1. *For each* $(\varphi,f) \in W^{1,p}(\mathbf{R}^-;\mathbf{R}^n;\eta) \times L^p(\mathbf{R}^+;\mathbf{R}^n;\eta)$, *define* $D(\varphi,f) = (\varphi, f + N(\varphi) + M(\varphi'))$, *and let* T *and* S *be the semigroups in Theorems* 2.1 *and* 3.6. *Then* D *maps* $W^{1,p}(\mathbf{R}^-;\mathbf{R}^n;\eta) \times L^p(\mathbf{R}^+;\mathbf{R}^n;\eta)$ *one-to-one and continuously onto itself, its inverse is the operator which maps* (φ,f) *into* $(\varphi, f - N(\varphi) - M(\varphi'))$, *and* $S(t) = DT(t)D^{-1}$, $T(t) = D^{-1}S(t)D$ *for all* $t \geqq 0$.

THEOREM 5.2. *For each* $(\varphi,f) \in L^p(\mathbf{R}^-;\mathbf{R}^n;\eta) \times W^{-1,p}(\mathbf{R}^+,\mathbf{R}^n;\eta)$, *define* $D(\varphi,f) + N(\varphi) + [M(\varphi)]')$ *and let* T *and* S *be the semigroups in Theorems* 2.4 *and* 4.5. *Then* D *maps* $L^p(\mathbf{R}^-;\mathbf{R}^n;\eta) \times W^{-1,p}(\mathbf{R}^+;\mathbf{R}^n;\eta)$ *one-to-one and continuously onto itself, its inverse is the operator which maps* (φ,f) *into* $(\varphi, f - N(\varphi) - [M(\varphi)]')$, *and* $S(t) = DT(t)D^{-1}$, $T(t) = D^{-1}S(t)D$ *for all* $t \geqq 0$.

THEOREM 5.3. *For each* $(\varphi,\zeta,f,g)\in L^p(\mathbf{R}^-;\mathbf{R}^n;\eta)\times\mathbf{R}^n\times L^p(\mathbf{R}^+;\mathbf{R}^n;\eta)\times L^p(\mathbf{R}^+;\mathbf{R}^n;\eta)$, *define* $D(\varphi,\zeta,f,g)=(\varphi,\zeta,f+N(\varphi),g+M(\varphi))$, *and let* T *and* S *be the semigroups in Theorems* 2.2 *and* 4.4. *Then* D *maps* $L^p(\mathbf{R}^-;\mathbf{R}^n;\eta)\times\mathbf{R}^n\times L^p(\mathbf{R}^+;\mathbf{R}^n;\eta)\times L^p(\mathbf{R}^+;\mathbf{R}^n;\eta)$ *one-to-one and continuously onto itself, its inverse is the operator which maps* (φ,ζ,f,g) *into* $(\varphi,\zeta,f-N(\varphi),g-M(\varphi))$, *and* $S(t)=DT(t)D^{-1}$, $T(t)=D^{-1}S(t)D$ *for all* $t\geqq 0$.

THEOREM 5.4. *For each* $(\varphi,\psi,\xi,f)\in L^p(\mathbf{R}^-;\mathbf{R}^n;\eta)\times L^p(\mathbf{R}^+;\mathbf{R}^n;\eta)\times\mathbf{R}^n\times L^p(\mathbf{R}^+;\mathbf{R}^n;\eta)$, *define* $D(\varphi,\psi,\xi,f)=(\varphi,\psi,\xi,f+N(\varphi)+M(\psi))$, *and let* T *and* S *be the semigroups in Theorems* 4.1 *and* 3.4. *Then* D *maps* $L^p(\mathbf{R}^-;\mathbf{R}^n;\eta)\times L^p(\mathbf{R}^+;\mathbf{R}^n;\eta)\times\mathbf{R}^n\times L^p(\mathbf{R}^+;\mathbf{R}^n;\eta)$ *one-to-one and continuously onto itself, its inverse is the operator which maps* (φ,ψ,ξ,f) *into* $(\varphi,\psi,\xi,f-N(\varphi)+M(\psi))$, *and* $S(t)=DT(t)D^{-1}$, $T(t)=D^{-1}S(t)D$ *for all* $t\geqq 0$.

THEOREM 5.5. *Suppose that the equation* (1.1) *is retarded, i.e. that* $\mu\equiv 0$. *For each* $(\varphi,\zeta,f)\in L^p(\mathbf{R}^-;\mathbf{R}^n;\eta)\times\mathbf{R}^n\times L^p(\mathbf{R}^+;\mathbf{R}^n;\eta)$, *define* $D(\varphi,\zeta,f)=(\varphi,\zeta,f+N(\varphi))$, *and let* T *and* S *be the semigroups in Theorems* 2.3 *and* 3.5. *Then* D *maps* $L^p(\mathbf{R}^-;\mathbf{R}^n;\eta)\times\mathbf{R}^n\times L^p(\mathbf{R}^+;\mathbf{R}^n;\eta)$ *one-to-one and continuously onto itself, its inverse is the operator which maps* (φ,ζ,f) *into* $(\varphi,\zeta,f-N(\varphi))$, *and* $S(t)=DT(t)D^{-1}$, $T(t)=D^{-1}S(t)D$ *for all* $t\geqq 0$.

All these theorems are direct consequences of the previous constructions.

There is also another connection between some of the semigroups presented above, which should be pointed out. One can get the two semigroups in $W^{1,p}(\mathbf{R}^-;\mathbf{R}^n;\eta)\times L^p(\mathbf{R}^+;\mathbf{R}^n;\eta)$ (those in Theorem 5.1) by restricting the semigroups in $L^p(\mathbf{R}^-;\mathbf{R}^n;\eta)\times W^{-1,p}(\mathbf{R}^+;\mathbf{R}^n;\eta)$ (those in Theorem 5.2) to the domains of their generators. In [27] this relationship plays a key role.

REFERENCES

[1] V. BARBU AND S. I. GROSSMAN, *Asymptotic behavior of linear integrodifferential systems*, Trans. Amer. Math. Soc., 173 (1972), pp. 277–288.

[2] C. BERNIER AND A. MANITIUS, *On semigroups in* $\mathbf{R}^n\times L^p$ *corresponding to differential equations with delays*, Canad. J. Math., XXX (1978), pp. 897–914.

[3] J. A. BURNS AND T. L. HERDMAN, *Adjoint semigroup theory for a class of functional differential equations*, this Journal, 7 (1976), pp. 729–745.

[4] J. A. BURNS, T. HERDMAN AND H. W. STECH, *Linear functional differential equations as semigroups on product spaces*, this Journal, 14 (1983), pp. 98–116.

[5] M. C. DELFOUR, *Status of the state space theory of linear hereditary differential systems with delays in state and control variables*, in Analysis and Optimization of Systems, A. Bensoussan and J. L. Lions, eds., Springer-Verlag, Berlin, 1980, pp. 83–96.

[6] M. C. DELFOUR AND A. MANITIUS, *The structural operator F and its role in the theory of retarded systems*, I, II, J. Math. Anal. Appl., 73 (1980), pp. 466–490; 74 (1980), pp. 359–381.

[7] O. DIEKMANN, *Volterra integral equations and semigroups of operators*, Report TW 197/80, Stichting Mathematisch Centrum, Amsterdam, 1980.

[8] O. DIEKMANN AND S. A. VAN GILS, *Invariant manifolds of Volterra integral equations of convolution type*, J. Differential Equations, to appear.

[9] S. A. VAN GILS, *Some studies in dynamical system theory*: I *Volterra integral equations of convolution type*, II *Hopf bifurcation and symmetry*, Ph. D. Thesis, Technische Hogeschool Delft, Delft, 1984.

[10] J. K. HALE, *Theory of Functional Differential Equations*, Springer-Verlag, Berlin, 1977.

[11] ———, *Functional differential equations with infinite delays*, J. Math. Anal. Appl., 48 (1974), pp. 276–383.

[12] J. K. HALE AND J. KATO, *Phase space for retarded equations with infinite delay*, Funkcial. Ekvac., 21 (1978), pp. 11–41.

[13] J. K. HALE AND K. R. MEYER, *A Class of Functional Equations of Neutral Type*, Mem. Amer. Math. Soc., 76, 1967.

[14] D. HENRY, *The adjoint of a linear functional differential equation and boundary value problems*, J. Differential Equations, 9 (1971), pp. 55–66.

[15] _____, *Linear autonomous neutral functional differential equations*, J. Differential Equations, 15 (1974), pp. 106–128.

[16] G. S. JORDAN, O. J. STAFFANS AND R. L. WHEELER, *Local analyticity in weighted L^1-spaces and applications to stability problems for Volterra equations*, Trans. Amer. Math. Soc., 274 (1982), pp. 749–782.

[17] F. KAPPEL, *Laplace-transform methods and linear autonomous functional-differential equations*, Berichte der Mathematisch-statistischen Sektion im Forschungszentrum Graz, Report, 64 (1976), pp. 1–62.

[18] _____, *Linear autonomous functional differential equations in the state space C^**, Technical Report 34, Technische Universität Graz, Graz, 1984.

[19] A. MANITIUS, *Completeness and F-completeness of eigenfunctions associated with retarded functional differential equations*, J. Differential Equations, 35 (1980), pp. 1–29.

[20] R. K. MILLER, *Nonlinear Volterra Integral Equations*, Benjamin, Menlo Park, CA, 1971.

[21] _____, *Linear Volterra integrodifferential equations as semigroups*, Funkcial. Ekvac., 17 (1974), pp. 39–55.

[22] R. K. MILLER AND G. R. SELL, *Volterra Integral Equations and Topological Dynamics*, Mem. Amer. Math. Soc., 102, 1970.

[23] T. NAITO, *Adjoint equations of autonomous linear functional differential equations with infinite retardations*, Tohoku Math. J., 28 (1976), pp. 135–143.

[24] _____, *On autonomous linear functional differential equations with infinite retardations*, J. Differential Equations, 21 (1976), pp. 297–315.

[25] _____, *On linear autonomous retarded equations with an abstract phase space for infinite delay*, J. Differential Equations, 33 (1979), pp. 74–91.

[26] R. S. PHILLIPS, *The adjoint semi-group*, Pacific J. Math. 5 (1955), pp. 269–283.

[27] D. SALAMON, *Control and Observation of Neutral Systems*, Research Notes in Mathematics 91, Pitman, London, 1984.

[28] O. J. STAFFANS, *On a neutral functional differential equation in a fading memory space*, J. Differential Equations, 50 (1983), pp. 183–217.

[29] _____, *A neutral* FDE *with stable D-operator is retarded*, J. Differential Equations, 49 (1983), pp. 208–217.

[30] _____, *The null space and the range of a convolution operator in a fading memory space*, Trans. Amer. Math. Soc., 281 (1984), pp. 361–388.

[31] _____, *Some well posed functional equations which generate semigroups*, J. Differential Equations, to appear.

[32] _____, *Semigroups generated by a convolution equation*, in Infinite Dimensional Systems, Proceedings, Retzhof, 1983, F. Kappel and W. Schappacher, eds., Springer-Verlag, Berlin, 1984.

[33] _____, *Extended initial and forcing function semigroups generated by a functional equation*, this Journal, 16 (1985), pp. 1034–1048.

[34] H. W. STECH, *On the adjoint theory for autonomous linear functional differential equations with unbounded delays*, J. Differential Equations, 27 (1978), pp. 421–443.

SIAM J. MATH. ANAL.
Vol. 17, No. 1, January 1986

007

STABILITY IN A REACTION-DIFFUSION MODEL OF MUTUALISM*

V. HUTSON†

Abstract. A reaction-diffusion model for the mutualistic interaction of two species is studied. A condition for the dominance of an equilibrium point in the bistable case is obtained, generalizing results for the well-known scalar case. It is also shown that the "hair trigger effect" operates when the corresponding kinetic system has a single globally asymptotically stable interior equilibrium point.

1. Introduction. Consider the pair of reaction-diffusion equations

$$u_{it}=u_i m_i(u)+\mu_i u_{ixx} \qquad (i=1,2) \tag{1.1}$$

with $u=(u_1,u_2)$, where the spatial region Ω is either $\mathbb{R}$, or a bounded open interval, a homogeneous Neumann condition being imposed on $\partial\Omega$. This system has been much studied recently, one of the most interesting applications being to biological problems with u interpreted as, say, population density. However, the mutualist case, that is when $\partial m_1/\partial u_2$, $\partial m_2/\partial u_1<0$, has been relatively neglected in the literature, and it is this case that is treated here, attention being concentrated on two of the most basic points of interest.

The first question concerns the dominance of an equilibrium point in the bistable case. To illustrate, take $\Omega=\mathbb{R}$. Recall that the scalar equation

$$u_t=um(u)+\mu u_{xx} \tag{1.2}$$

is said to be bistable when the corresponding kinetic equation $\dot{u}=um(u)$ has exactly three equilibrium points 0, $\tilde{u}$ and u^* with $0<\tilde{u}<u^*$, the points 0 and u^* being asymptotically stable and $\tilde{u}$ unstable. For (1.2), the dominance of one of the asymptotically stable equilibrium points over the other has been extensively discussed, see for example [1] and [4]. When u^* is dominant, if the initial value $u(x,0)$ is not too small on a sufficiently large (but finite) interval, then $u(x,t)\to u^*$ uniformly on compact sets as $t\to\infty$. For the two species mutualist case the phase plane may have the form shown in Fig. 1a, and by analogy this will be called the two species bistable case. We derive in §3 a condition for this case which will ensure that u^* is dominant.

Secondly, we consider in §4 the case when the kinetic system corresponding to (1.1) has a single globally asymptotically stable interior equilibrium point Q. If the origin is a source, it is shown in [5] that all solutions with initial values neither of whose components are identically zero are attracted to Q (the hair trigger effect). In §4 we extend this result to the case when the origin is a saddle point.

Our results have applications to the resolution of a paradox concerning the distribution of mutualistic systems in nature; for general biological background on mutualism see [2], [8], [9], [10], [11], [13] and [14]. If the kinetic model for mutualism is adopted, it is clear from Fig. 1a that the occurrence of obligate/obligate mutualisms[1] is

* Received by the editors December 13, 1983, and in revised form November 16, 1984.
† Department of Applied Mathematics, The University, Sheffield S10 2TN, England.

[1] The terminology of [2] is used: mutualism is an interaction between species that is beneficial to both; an obligate and a facultative mutualist cannot (respectively can) survive without its partner; on the other hand symbiosis is the living together of two organisms in close association.

relatively unlikely in an unstable environment. For if a natural disaster should sweep the population vector into the hatched region, the extinction of both species will follow. However, such mutualisms are common even in very variable temperate and arctic regions, see [8]. If the species may diffuse, the results of §3 show that the system is a great deal more resilient, for if some, perhaps quite limited region is left relatively unaffected, the species will return to their coexistence levels u^*. That is, the domain of attraction of Q is much larger for the reaction-diffusion system than for the kinetic system.

2. Preliminaries. We consider the problem of finding classical solutions of (1.1) given that the initial value $u(x,0)$ is bounded nonnegative and uniformly continuous on $\bar{\Omega}$, and that a homogeneous Neumann condition holds on $\partial\Omega$ when Ω is finite.

It will be assumed throughout that the following conditions are satisfied.

(C1) $\mu_1,\ \mu_2 > 0$.

(C2) $m_1,\ m_2 \in C^1(\mathbb{R}_2^+, \mathbb{R})$, and for all $u_1,\ u_2 \geqq 0$,

(i) $\partial m_1/\partial u_1,\ \partial m_2/\partial u_2 < 0$;

(ii) $\partial m_1/\partial u_2,\ \partial m_2/\partial u_1 > 0$.

(C3) There exist $\alpha_1,\ \alpha_2 \in \mathbb{R}$ such that

$$m_1(u_1,u_2) < 0 \qquad (u_1 \geqq \alpha_1, u_2 \geqq 0),$$
$$m_2(u_1,u_2) < 0 \qquad (u_1 \geqq 0, u_2 \geqq \alpha_2).$$

(C2)(i) implies that, because of intraspecific competition, the per capita rate of increase of each species decreases as its population increases, while (C2)(ii) is the mutualist assumption that an increase in the population of one species increases the per capita growth rate of the other. (C3) requires further that intraspecific competition causes the population density of each species to decrease if it is large enough no matter how numerous the other species may be; this appears essential if the equations are to reflect biological reality, and corresponds to a "finite world" assumption—for further discussion see [10], [11]. Of course a Lotka–Volterra system cannot satisfy both (C2) and (C3). Finally, for the equilibrium points of (1.1), assume that one of the following holds.

($C4_1$) There are in the interior of $\mathbb{R}_2^+$ exactly two equilibrium points $\tilde{u} = (a_1, a_2)$ and $u^* = (A_1, A_2)$ where $a_1 < A_1$, $a_2 < A_2$ (Fig. 1a).

($C4_2$) There is exactly one equilibrium point $u^* = (A_1, A_2)$ in the interior of $\mathbb{R}_2^+$ (Fig. 1b). When ($C4_1$) holds, apart from the configuration shown in Fig. 1a, there are other possibilities, for example that u_1 is facultative and u_2 is obligate, when the u_1 isocline cuts the positive u_1 axis at $(a,0)$ say, in which case $(0,0)$ will be a saddle point and $(a,0)$ will be asymptotically stable. When ($C4_2$) holds, apart from Fig. 1b, there are again other possibilities; for example u_2 may also be facultative, in which case the u_2 isocline will cut the positive u_2 axis.

For later reference some properties of the solutions of the kinetic equation are now quoted from [7].

PROPOSITION 2.1. *In case* ($C4_1$) *there is a one-dimensional stable manifold S through* $\tilde{u}$. *The point* u^* *is a global attractor for all population vectors whose initial values lie to the right of S, while if the initial values lie to the left of S* (*that is in the hatched region in Fig.* 1a), *at least one population tends to zero as* $t \to \infty$.

In case ($C4_2$), u^* *is globally asymptotically stable* (*with respect to all populations with* $u_1(0),\ u_2(0) > 0$).

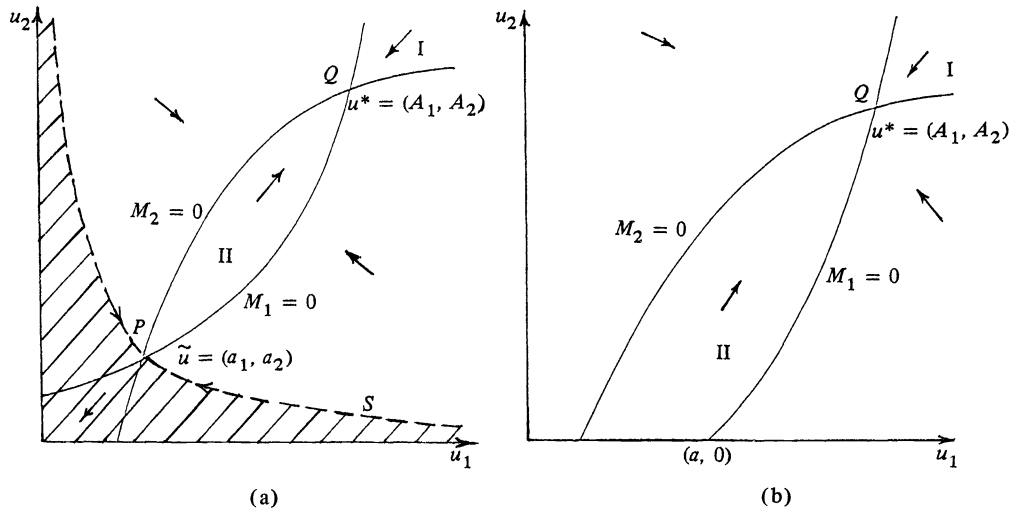


FIG. 1. *Possible configurations for the zero isoclines of cases* $(C4_1)$ *and* $(C4_2)$ *respectively, other possibilities being described in the text. In Fig.* 1a *both species are obligate, while in Fig.* 1b *the first species is obligate and the second facultative.*

Sub and supersolution techniques are particularly effective for mutualistic interactions. The basic mathematical apparatus is given in [5] and [6], and for later reference we outline the salient results specialised to (1.1). To fit the framework of [5], set $M_i(u)=u_i m_i$ and extend M_1 and M_2 to $\mathbb{R}_2$ by setting $M_1(u_1,u_2)=M_1(u_1,0)$, $M_2(u_1,u_2)=0$ for $u_1>0$ and $u_2<0$, $M_1(u_1,u_2)=M_2(u_1,u_2)=0$ for u_1, $u_2\leqq 0$, and $M_1(u_1,u_2)=0$, $M_2(u_1,u_2)=M_2(0,u_2)$ for $u_1<0$ and $u_2>0$. With $u=(u_1,u_2)$, $v=(v_1,v_2)$ the relation $u<v$ (respectively $u\leqq v$) means that $u_i<v_i$ (respectively $u_i\leqq v_i$) for $i=1,2$.

If $\Omega=\mathbb{R}$, a regular subsolution $\underline{u}$ is defined on an open set $X\subset\mathbb{R}\times\mathbb{R}$, and must satisfy

$$\underline{u}_{it}\leqq M_i(\underline{u})+\mu_i\underline{u}_{ixx} \qquad (i=1,2)$$

on X, with all derivatives appearing being continuous. Let $X=\Omega\times(0,\tau)$ for some $\tau>0$. $\underline{u}$ is said to be a subsolution if there is an $\varepsilon>0$ such that for every point $P\in\overline{X}$ there is a finite collection $\{\underline{u}^{(1)},\cdots,\underline{u}^{(\alpha)}\}$ of regular subsolutions in $B(P,\varepsilon)$, the ε-ball centre P, such that in $\overline{X}\cap B(P,\varepsilon)$,

$$\underline{u}(x,t)=\max_{1\leqq k\leqq\alpha}\underline{u}^{(k)}(x,t),$$

the max being taken componentwise. If Ω is bounded, it is required in addition that the outward normal derivative $\partial\underline{u}/\partial\nu\leqq 0$ on $\partial\Omega$. Supersolutions are defined analogously by reversing the inequalities and substituting "min" for "max". Of course every solution is a sub and supersolution.

PROPOSITION 2.2. *Under the stated conditions on* $u(x,0)$ *the following hold.*

(i) *Every solution of* (1.1) *is nonnegative and satisfies* $u(x,t)\leqq M$ *for some* M *and all* $t>0$. *Further, if* α *lies in a region* I (*Fig.* 1), *then* $u(x,t)\leqq\alpha$ *for large enough* t.

(ii) *A unique classical solution exists and is uniformly continuous in* $\overline{X}$.

(iii) *If neither of* $u_1(x,0)$, $u_2(x,0)$ *is identically zero, then* $u(x,t)>0$ *for all* $x\in\Omega$, $t>0$.

Proof. (i) The bound on u follows from [4, Thm. 5.1] when the properties of the kinetic system given by Proposition 2.1 are used. (ii) is a consequence of [4, Thm. 3.1] and [12, Thm. 14.4]. (iii) This follows from the strong maximum principle, see [12].

THEOREM 2.3. *Let $\underline{u}$ and $\bar{u}$ be bounded uniformly continuous sub and supersolutions in $\bar{X}$. If $\underline{u}(x,0) \leqq \bar{u}(x,0)$ then $\underline{u}(x,t) \leqq \bar{u}(x,t)$ in X.*

THEOREM 2.4. *Assume that u, $\bar{u}$ are bounded uniformly continuous stationary sub and supersolutions respectively with $\underline{u} \leqq \bar{u}$. Let u be the solution with $u(x,0) = \underline{u}(x)$. Then there is a stationary solution w with $\underline{u} \leqq w \leqq \bar{u}$ such that as $t \to \infty$, $u(x,t) \to w(x)$ uniformly on compact sets.*

Let $\Omega = \mathbb{R}$, and suppose that there is a neighborhood N of 0 such that $w(x+y) \geqq \underline{u}(x)$ for all $y \in N$. Then w is a constant.

3. Dominance for a two species bistable system. It will be assumed throughout this section that ($C4_1$) holds, a possible configuration of the isoclines being given in Fig. 1a. Theorem 3.1 shows that condition (D_Q) below yields an analogue of dominance for a one species problem.

Define $\mu = (\mu_1 \mu_2)^{1/2}$, and set

$$
\begin{aligned}
M(u) &= \mu \min\left[\mu_1^{-1} M_1(u,u), \mu_2^{-1} M_2(u,u)\right], \\
V(u) &= \int_0^u M(s)\, ds.
\end{aligned} \tag{3.1}
$$

$M(u)$ is negative for small and large $u > 0$, and it follows from the definition that if $\alpha > 0$ is a zero of M, then (α, α) lies in region II, whence a_1, $a_2 \leqq \alpha \leqq A_1$, A_2. Assume that the following holds (Fig. 2).

(D_Q) There exists a number $b > 0$ such that $V(u) < V(b)$ for $0 \leqq u < b$.

Note that a rescaling of u_1, u_2 in (1.1) may be advantageous in arranging for this condition to hold. The idea is roughly that as much as possible of the line $u_1 = u_2$ should lie in region II in Fig. 1a.

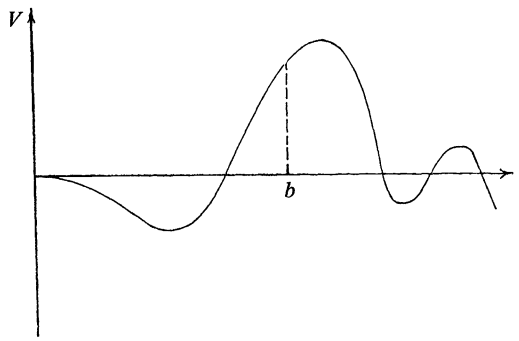


FIG. 2. *V when condition (D_Q) holds.*

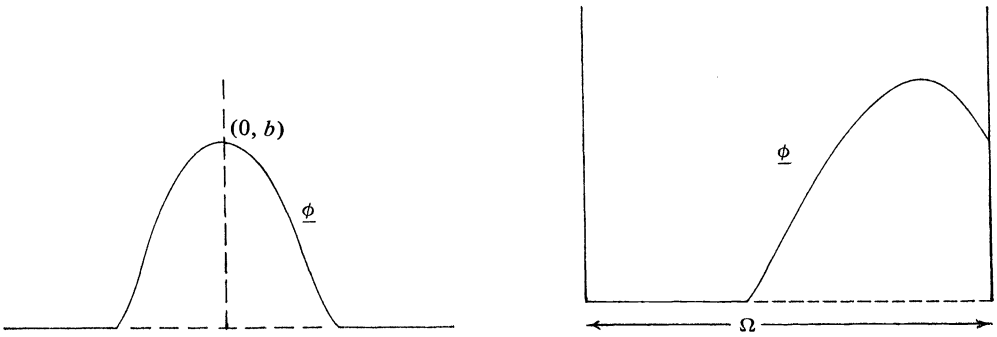
Assume that $0 \in \bar{\Omega}$, and let ϕ be the solution of the equation

$$\mu \phi'' + M(\phi) = 0, \tag{3.2}$$

with $\phi(0) = b$, $\phi'(0) = 0$. From the first integral

$$\mu \left[\phi'(x)\right]^2 = 2\left[V(b) - V(\phi(x))\right], \tag{3.3}$$

it is clear that if $\Omega=\mathbb{R}$, ϕ will eventually reach zero at points symmetrical with respect to 0. Define $\underline{\phi}(x)=\max[0,\phi(x)]$, see Fig. 3a. If Ω is finite, ϕ may first reach the boundary, where from (3.3) $\partial\phi/\partial\nu<0$ (Fig. 3b).



(a) (b)

FIG. 3. *Possibilities for $\underline{\phi}$ when $\Omega=\mathbb{R}$ and Ω is finite respectively.*

With $\underline{\Phi}=(\underline{\phi},\underline{\phi})$, we have the following result for the system (1.1).

THEOREM 3.1. *Suppose that $0\in\bar{\Omega}$, and assume that conditions* (C1)–(C3), ($C4_1$) *and* (D_Q) *hold. Let $u(x,t)$ be the solution of* (1.1) *with $u(x,0)\geqq\underline{\Phi}$ on Ω. Then $u\to u^*$ uniformly on compact sets as $t\to\infty$.*

Proof. Note first that $\underline{\Phi}$ is a subsolution of (1.2). For by (3.1) and (3.2),

$$\mu_i\underline{\phi}_{xx}+M_i(\underline{\phi},\underline{\phi})\geqq\mu_i\underline{\phi}_{xx}+\mu_i\mu^{-1}M(\underline{\phi})=0,$$

and $(0,0)$ is a subsolution. For $\Omega=\mathbb{R}$ the assertion follows as $\underline{\Phi}$ is the local maximum of regular subsolutions, and for Ω bounded also $\partial\underline{\Phi}/\partial\nu\leqq 0$ on $\partial\Omega$.

Choose any point $(\bar{u}_1,\bar{u}_2)$ in region I with $\bar{u}_i\geqq\sup_{x\in\Omega}u_i(x,0)$ for $i=1,2$. Let $\bar{u}$ be the (spatially independent) solution of (1.2) with $\bar{u}(0)=(\bar{u}_1,\bar{u}_2)$. Define $\underline{u}(x,t)$ to be the solution of (1.1) with $\underline{u}(x,0)=\underline{\Phi}$. Then by Theorem 2.3, for all $t>0$,

$$\underline{\Phi}(x)\leqq\underline{u}(x,t)\leqq u(x,t)\leqq\bar{u}(t).$$

Now from Proposition 2.1, $\lim_{t\to\infty}\bar{u}(t)=u^*$. Furthermore, $\underline{\Phi}$ and u^* are respectively stationary sub and supersolutions with $\underline{\Phi}\leqq u^*$. Hence from Theorem 2.4, there is a stationary solution w with $\underline{\Phi}\leqq w\leqq u^*$ such that $\underline{u}(x,t)\to w(x)$ uniformly on compact sets as $t\to\infty$. Therefore the proof will be complete if it can be shown that $w=u^*$.

Suppose first that $\Omega=\mathbb{R}$. Then $w(x)>\underline{\Phi}(x)$ for $x\in\Omega$. For if not there is an x_0 and an $i=1$ or 2 such that $w_i(x_0)=\underline{\phi}(x_0)$. But $\underline{\Phi}$ is a subsolution, $w(x)$ is a regular supersolution and $\underline{\Phi}\leqq w$. Hence from [5, Thm. 1], $w_i(x)=\underline{\phi}(x)$ for all x. This is impossible as $\underline{\phi}(x)$ has a discontinuous first derivative.

Now $\underline{\phi}$ has compact support. Hence if Ω_0 is any compact set containing the support of $\underline{\phi}$ in its interior, there exists $\delta=(\delta_1,\delta_2)>0$ such that $w(x)\geqq\underline{\Phi}(x)+\delta$ for $x\in\Omega_0$. It follows from uniform continuity that there is a neighborhood N of zero such that $w(x+y)>\underline{\Phi}(x)$ for all $y\in N$ and $x\in\mathbb{R}$. Therefore by Theorem 2.4, w is a constant.

Finally, if $\alpha>0$ is any zero of M, the point (α,α) lies in region II (Fig. 1a). But from the construction of $\underline{\phi}$, $\underline{\phi}(0)$ and so w_1 is greater than the smallest positive zero of

M. Therefore $w \neq \tilde{u}$, and so $w = u^*$. This completes the proof for $\Omega = \mathbb{R}$. It follows immediately that the only stationary solution w of (1.1) satisfying $\underline{\Phi} \leqq w \leqq u^*$ is u^* itself, a fact which will be used below.

When Ω is bounded, the result will be proved by first constructing an extension of w to $\mathbb{R}$. Consider then the stationary solution w of (1.1) on $\bar{\Omega}$ satisfying $\partial w/\partial \nu = 0$ on $\partial\Omega$. Let $\Omega = (\xi, \xi + \gamma)$ where $\xi \leqq 0 \leqq \xi + \gamma$. Define an extension of w to $\mathbb{R}$ by successively reflecting w in the lines $x = \xi$, $\xi - \gamma, \cdots$, and $x = \xi + \gamma$, $\xi + 2\gamma, \cdots$, and note that w then has period 2γ. Because $\partial w/\partial x = 0$ on $x = \xi$ and $x = \xi + \gamma$, w is a C^2-function which is a solution of (1.1) for all x. It is also clear from the evenness of $\underline{\phi}$ and the fact that it is decreasing for $x > 0$ that $w \geqq \underline{\Phi}$. This shows that w extended in this manner is a solution of (1.1) on $\mathbb{R}$ satisfying $\underline{\Phi} \leqq w \leqq u^*$, so by the final remark of the preceding paragraph, $w = u^*$. This completes the proof.

A similar mutualist system has been considered in [3] although from a different point of view. There Ω must be bounded, and a condition is imposed which requires among other things that μ_1 and μ_2 are sufficiently large relative to the length of Ω. Then u tends to a spatially independent solution for large t, so that the asymptotic behavior of u may be deduced from that of the solution of the analogous kinetic system. Under certain further restrictions it will be the case that if the space average of $u(x, 0)$ lies to the right of S, the solution will eventually return towards Q. The conditions in [3] and Theorem 3.1 are of course of quite a different nature, but from the point of view of biological applications the principal point of interest in Theorem 3.1 is that it shows that the restriction on the space average of $u(x, 0)$ and on μ_1, μ_2 can be removed if D_Q is imposed.

4. Global asymptotic stability. If there is exactly one interior equilibrium point for the kinetic equations, this point is globally asymptotically stable, and the aim now is to show that an analogous property also holds when there is diffusion. This extends a result in [5, §4] to the case where there is a saddle point at the origin.

THEOREM 4.1. *Assume that conditions* (C1)–(C3) *and* ($C4_2$) *hold. Let* $u(x, t)$ *be the solution of* (1.1) *with* $u(x, 0) \geqq 0$ *on* Ω, *and suppose that neither* $u_1(x, 0)$ *nor* $u_2(x, 0)$ *is identically zero. Then as* $t \to \infty$, $u \to u^*$ *uniformly on compact sets.*

Proof. By Proposition 2.2, $u(x, t) > 0$ for $t > 0$. Therefore, by shifting to a new time origin we may assume that $u(x, t) > 0$ in $\bar{\Omega}$ for $t \geqq 0$.

(i) Ω bounded. The proof is very simple. Let $\underline{u}(t)$ and $\bar{u}(t)$ be solutions of (2.1) with

$$\underline{u}_i(0) = \min_{x \in \bar{\Omega}} u_i(x, 0), \qquad \bar{u}_i(0) = \max_{x \in \bar{\Omega}} u_i(x, 0)$$

for $i = 1, 2$; by the remark above $\underline{u}(0) > 0$. Now $\underline{u}$, $\bar{u}$ are sub and supersolutions respectively, so by Theorem 2.3, $\underline{u}(t) \leqq u(x, t) \leqq \bar{u}(t)$ for $t > 0$. The result follows, for by Proposition 2.1, $\lim_{t \to \infty} \underline{u}(t) = \lim_{t \to \infty} \bar{u}(t) = u^*$.

(ii) $\Omega = \mathbb{R}$. A more elaborate proof is needed since it is not necessarily true that $\underline{u}(0) > 0$. If both species are facultative, then the origin lies in region II and is a source for the kinetic problem, so the argument of [5, §4] yields the result. However, if one species is obligate and one facultative, as is the case in Fig. 1, this argument does not work.

So far as the upper bound of u is concerned, of course it follows as in (i) above $u(x, t) \leqq \bar{u}(t)$, where $\lim_{t \to \infty} \bar{u}(t) = u^*$. We shall show that there is a function $z(x, t)$ which approaches u^* uniformly on compact sets as $t \to \infty$ such that $u(x, t) \geqq z(x, t)$ for large enough t, from which the assertion of the theorem follows.

From the assumption on $M_1(v_1,0)=v_1m_1(v_1,0)$ there exists an a_0 with $0<a_0<a$ and a $\gamma>0$ such that $M_1(v_1,0)\geqq\gamma^2\mu_1v_1$ for $0\leqq v_1\leqq a_0$. Since $u_1(x,0)>0$ for $x\in\Omega$, there is a δ with $0<\delta<a_0$ such that

$$\delta<\min_{-l_1\leqq x\leqq l_1}u_1(x,0),$$

where $l_1=\pi/2\gamma$. Set

$$v_1(x)=\begin{cases}\delta\cos\gamma x & (|x|\leqq\pi/2\gamma),\\ 0 & (|x|>\pi/2\gamma),\end{cases}$$

and let $v_1(x,t)$ be the solution of the one-dimensional problem

$$v_{1t}=M_1(v_1,0)+\mu_1v_{1xx}, \tag{4.1}$$

with $v_1(x,0)=v_1(x)$. Put $v(x,t)=(v_1(x,t),0)$ and note that v is a solution of (1.2).

LEMMA 4.2. *$v_1(x)$ is a stationary subsolution of* (4.1), *and the following hold.*

(i) $v_1(x)\leqq v_1(x,t)\leqq a$ $(t\geqq0)$.

(ii) $\partial v_1(x,t)/\partial t\geqq0$ *for* $x\in\Omega$ *and* $t>0$.

(iii) $v_1(x,t)\to$ *a uniformly on compact sets as* $t\to\infty$.

(iv) $u(x,t)\geqq v(x,t)$ *for* $t\geqq0$.

Proof. We have

$$M_1(v_1,0)+\mu_1v_{1xx}\geqq\gamma^2\mu_1v_1+\mu_1v_{1xx}=0,$$

and 0 is a solution. Hence v_1 is the local maximum of regular subsolutions, and is therefore a subsolution of (4.1). Also $a>v_1(x)$ is a supersolution, so (i), (ii) and (iii) follow from Theorems 4.1, 4.2, 4.8 respectively of [4]. Finally, $v(x,t)$ is a solution of (1.1), and since $u(x,0)\geqq v(x,0)$, Theorem 2.3 yields (iv). This concludes the proof of the lemma.

The lemma shows that on any compact set the lower bound of u_1 moves into region II for large t. Since $u_2>0$ it is plausible from the direction of the vector field that u is swept towards u^*. This is now proved by constructing another subsolution whose first component is based on the value of $v_1(x,t)$ for large t.

From the assumptions on $M_2(z_1,z_2)=z_2m_2(z_1,z_2)$ there are numbers $\varepsilon,b,\rho>0$ such that $M_2(a-\varepsilon,z_2)\geqq\rho^2\mu_2z_2$ for $0\leqq z_2\leqq b$. From (iii) of Lemma 4.2, there exists a t_0 such that $v_1(x,t_0)\geqq a-\varepsilon$ for $|x|\leqq\pi/2\rho$. Since $u_2(x,t_0)>0$ for $x\in\Omega$, there is a $\sigma>0$ such that

$$\sigma<\min_{-l_2\leqq x\leqq l_2}u_2(x,t_0),$$

where $l_2=\pi/2\rho$. Define $z(x)=(z_1(x),z_2(x))$ where

$$z_1(x)=v_1(x,t_0)\qquad(x\in\Omega),$$

$$z_2(x)=\begin{cases}\sigma\cos\rho x & (|x|\leqq\pi/2\rho),\\ 0 & (|x|>\pi/2\rho),\end{cases}$$

and let $z(x,t)$ be the solution of (2.1) with $z(x,t_0)=z(x)$. From the above construction and Lemma 4.2(iv), $u(x,t_0)\geqq z(x,t_0)\geqq v(x,t_0)$, and since $u(x,t)$, $z(x,t)$ and $v(x,t)$

are solutions of (1.2), from Theorem 2.3, $u(x,t) \geqq z(x,t) \geqq v(x,t)$ for $t \geqq t_0$. Now

$$\begin{aligned} M_1(z_1(x), z_2(x)) + \mu_1 z_{1xx}(x) &\geqq M_1(z_1(x), 0) + \mu_1 z_{1xx}(x) \\ &= M_1(v_1(x,t_0), 0) + \mu_1 \frac{\partial^2}{\partial x^2} v_1(x,t_0) \\ &= \frac{\partial v_1}{\partial t_0}(x,t_0) \\ &\geqq 0, \end{aligned}$$

by Lemma 4.2(ii). Also,

$$\begin{aligned} M_2(z_1(x), z_2(x)) + \mu_2 z_{2xx}(x) &\geqq M_2(a-\varepsilon, z_2(x)) + \mu_2 z_{2xx}(x) \\ &\geqq \rho^2 \mu_2 z_2(x) + \mu_2 z_{2xx}(x) = 0, \end{aligned}$$

and the left-hand side is zero when $z_2 = 0$. Thus $z(x)$ is the local maximum of regular subsolution, and so is a subsolution.

Now $z(x)$ is a subsolution, u^* is a supersolution and $z \leqq u^*$, so it follows from Theorem 2.4 that there is a stationary solution $w(x)$ with $z(x) \leqq w(x) \leqq u^*$ such that $z(x,t) \to w(x)$ uniformly on compact sets as $t \to \infty$. The final step is to show that $w = u^*$.

As noted above, $z_1(x,t) \geqq v_1(x,t)$, and since $v_1(x,t) \to a$ uniformly on compact sets, it follows that $w_1(x) \geqq a \geqq z_1(x)$. Also, since z_2 has compact support, by an argument used in proving Theorem 3.1, there is an $\alpha > 0$ such that $w_2(x) \geqq z_2(x) + \alpha$ on the support of z_2. There is therefore a neighborhood N of 0 such that $w(x+y) \geqq z(x)$ for $y \in N$. Therefore, by Theorem 2.4, w is a constant. However, the only constant solution satisfying $w \geqq z(x)$ is u^*, so $w = u^*$. This concludes the proof.

Acknowledgment. I would like to express my gratitude to Dr. R. Law of the University of York for discussions concerning this problem.

REFERENCES

[1] D. G. Aronson and H. F. Weinberger, *Nonlinear diffusion in population genetics, combustion, and nerve propagation*, in Partial Differential Equations and Related Topics, Lecture Notes in Mathematics 446, Springer-Verlag, Berlin, 1975.

[2] D. H. Boucher, S. James and K. H. Keeler, *The ecology of mutualism*, Ann. Rev. Ecol. Syst., 13 (1982), pp. 315–347.

[3] E. Conway, D. Hoff and J. Smoller, *Large time behavior of systems of nonlinear reaction-diffusion equations*, SIAM J. Appl. Math., 35 (1978), pp. 1–16.

[4] P. C. Fife, *Mathematical Aspects of Reacting and Diffusing Systems*, Springer-Verlag, New York, 1979.

[5] P. C. Fife and M. M. Tang, *Comparison principles for reaction-diffusion systems: irregular comparison functions and applications to questions of stability and speed of propagation of disturbances*, J. Differential Equations, 40 (1981), pp. 165–185.

[6] ———, *Corrigenda: comparison principles for reaction-diffusion systems: irregular comparison functions and applications to questions of stability and speed of propagation of disturbances*, J. Differential Equations, 51 (1984), pp. 442–447.

[7] M. W. Hirsch and S. Smale, *Differential Equations, Dynamical Systems, and Linear Algebra*, Academic Press, New York, 1974.

[8] D. H. LEWIS, *Mutualist lives*, Nature, 297 (1982), p. 176.
[9] L. MARGULIS, *Symbiosis in Cell Evolution*, Freeman, San Francisco, 1981.
[10] R. M. MAY, *Models for two interacting populations*, in Theoretical Ecology: Principles and Applications, R. M. May, ed., Saunders, Philadelphia, 1981, pp. 94–100.
[11] _____, *Mutualistic interactions among species*, Nature, 296 (1982), pp. 803–804.
[12] J. SMOLLER, *Shock Waves and Reaction-Diffusion Equations*, Springer-Verlag, New York, 1982.
[13] P. J. VAN BENEDEN, *Les commensaux et les parasites*, Biblio. Sci. Int., Paris, 1875.
[14] J. H. VAN DER MEER AND D. H. BOUCHER, *Varieties of mutualistic interactions in population models*, J. Theor. Biol., 74 (1978), pp. 549–558.

SIAM J. MATH. ANAL.
Vol. 17, No. 1, January 1986

008

A PARABOLIC-HYPERBOLIC FREE BOUNDARY PROBLEM*

ANTONIO FASANO† AND MARIO PRIMICERIO†

Abstract. Change of phase problems with space-dependent melting temperature are considered. A weak formulation is given, pointing out that relevant differences occur with respect to the case of constant melting temperature. Next, analyzing the case in which weak solutions are suitably smooth, it is found that (i) mushy regions may appear even if volumetric heat sources are absent; (ii) a differential system is satisfied, consisting of parabolic equations and first order hyperbolic equations coupled through free boundaries; (iii) the interface conditions have the form of "unilateral constraints". Finally, a model problem is studied and the existence of a smooth solution is proved.

1. Introduction. In this paper we study a mathematical model for change of phase processes (e.g., freezing of lakes, thawing of glaciers, etc.) in which the melting temperature is space-dependent.

Among the main features of this model, we point out that "mushy" regions (see [1], [8], [11], [12], [13]) may appear even if volumetric heat sources are absent. Moreover, in such regions while temperature is locally constant (and equal to the melting temperature), heat conduction makes energy evolve with time.

This model could be the starting point of a classical theory for melting or solidification of alloys (see [2], [4], [14]). Actually, the results presented here show that the possible appearance of a mushy region during the solidification of an alloy is not necessarily related to the gap between the solidus and liquidus curves in the phase diagram. Indeed, when the concentration of the diluted component is not uniformly distributed and its diffusion is negligible, then an alloy in which the liquidus and solidus curves were supposed to be almost coincident would behave essentially as a medium with space-dependent melting temperature.

In §2 we will write the weak formulation of the problem in a general setting. In §3 we will sketch some properties of the classical solution in one space dimension. We will show that a differential system has to be satisfied, consisting of parabolic equations and first order hyperbolic equations coupled through free boundaries. Moroever, the interface conditions have the form of "unilateral constraints". In §4 a model problem will be considered, clearly exhibiting the spontaneous appearance of a mushy region. A well posedness theorem is stated and an outline of its proof is given. Details and a more comprehensive discussion of the model will appear in a forthcoming paper ([6]).

2. Weak formulation. Let x be a point in R^n and let t denote time, u temperature, k thermal conductivity, c heat capacity, r rate of heat supply per unit volume. Assume that the melting temperature u_m is a smooth function of x and that k, c, r, depend on the difference $w = u - u_m$ only, being bounded and smooth for $w \neq 0$. Moreover, $c, k \geqq a$ for some positive constant a.

The thermal energy E is defined as

$$E(x,t) = \int_0^{w(x,t)} c(y)\,dy + \Lambda \operatorname{sgn}^+ w(x,t), \tag{2.1}$$

*Received by the editors September 11, 1983, and in revised form June 25, 1984. This work was partially supported by the GNFM of the Italian CNR and by the University of Firenze.

†c/o Istituto Matematico "Ulisse Dini" viale Morgagni 67/a, I 50134 Firenze, Italy.

where Λ is the latent heat and

$$\operatorname{sgn}^+ z = \begin{cases} 0, & z<0, \\ [0,1], & z=0, \\ 1, & z>0. \end{cases}$$

Obviously, w can be found in terms of E as a single valued function $b(E)$, and

$$u(x,t)=u_m(x)+b(E(x,t)). \tag{2.2}$$

Moreover, we set

$$g(w)=\int_0^w k(y)\,dy, \tag{2.3}$$

$$B(E)=g(b(E)). \tag{2.4}$$

Given $T>0$ and a bounded domain $Q\subset R^n$ with smooth boundary ∂Q and outer normal n, let us define a weak solution of the change of phase problem with data $E_0(x)$, the initial energy, and $U(x,t)$, the boundary temperature. Here, we will assume that the subset of $\partial Q\times(0,T)$ where $U=u_m$ has zero measure.[1] We extend the functions

$$K(E)=k(b(E)), \qquad R(E)=r(b(E)), \qquad E\neq(0,\Lambda), \tag{2.5}$$

over the interval $(0,\Lambda)$ as linear interpolations.

We say that $E\in L^\infty(Q\times(0,T))$ is a weak solution if the following equation is satisfied for any $F\in C^\infty(\bar{Q}\times 0,T)$, vanishing for $t=T$ and on $\partial Q\times(0,T)$,

$$\int_0^T\int_Q [EF_t+B\operatorname{div}\operatorname{grad}F-K\operatorname{grad}u_m\cdot\operatorname{grad}F+RF]\,dx\,dt + \int_Q E_0(x)F(x,0)\,dx-\int_{\partial Q\times(0,T)} B_0\operatorname{grad}F\cdot n\,dS=0, \tag{2.6}$$

where $B_0=g(U-u_m)$.

3. Classical solutions. A solution $E(x,t)$ of the above mentioned problem is said to be a classical solution if it possesses some additional regularity properties which will be listed below for the case $n=1$, $Q=(0,1)$:

i) The function $u(x,t)$ given by (2.2) is continuous in the closure of $P=(0,1)\times(0,T)$.

ii) The sets

$$L=\{(x,t)\in P: u(x,t)>u_m\},$$
$$S=\{(x,t)\in P: u(x,t)<u_m\},$$
$$M=\mathring{N}, \qquad N=P\setminus(L\cup S),$$

are such that those parts of their boundaries which lie in P consist of N curves $x=s_i(t)$, $t\in I_i=(t_i',t_i'')\subset(0,T)$ such that $s_i\in C(\bar{I}_i)\cap C^1(I_i)$, $i=1,2,\cdots,N$.

iii) $u\in C^{2,1}(L\cup S)$ and u_x is continuous on each side of the curves $x=s_i(t)$.

iv) $E\in C^{1,1}(M)\cap C(\bar{M})$.

---

[1] When this condition is not fullfilled, the boundary data to be prescribed are the temperature or the energy, according to the thermal properties of the medium. This nontrivial point is discussed in [6], together with the analysis of the problem with Neumann data.

We will call L and S the liquid and the solid regions respectively, while M will be called the mushy region. In M it is $u(x,t)=u_m(x)$, and $E\in[0,\Lambda]$ (regions where $E=0$ or $E=\Lambda$ identically could be included in M as well).

It is immediately seen that the parabolic equation

$$cu_t=(ku_x)_x+r \tag{3.1}$$

holds in L and in S, and that the first order hyperbolic equation

$$E_t=\big(K(E)u'_m(x)\big)_x+R(E) \tag{3.2}$$

is satisfied in M. Moreover it is immediately found that the initial and boundary conditions are satisfied in a classical sense,

(3.3)

$$E(x,0)=E_0(x),\quad x\in(0,1),\qquad u(0,t)=U(0,t),\quad u(1,t)=U(1,t),\quad t\in(0,T).$$

To find the conditions satisfied on interfaces, multiply (3.1) and (3.2) by a test function F and use Green's theorem on each connected component of L, S and M. If we denote, for any function $f(x,t)$, $[f]_i=f(s_i(t)+,t)-f(s_i(t)-,t)$, we find by comparison with (2.6):

$$[E]_i\dot{s}_i+[Ku_x]_i=0,\qquad t\in I_i,\quad i=1,2,\cdots,N. \tag{3.4}$$

A careful analysis of condition (3.4) shows that the following local characteristic speed

$$v_0(x)=-hu'_m(x),\qquad h=\frac{dK}{dE},\quad E\in(0,\Lambda) \tag{3.5}$$

determines the form of the free boundary conditions. Namely:

(A) When $\dot{s}(t)>v_0(s(t))$ and $x=s(t)$ is a boundary between solid (for $x<s(t)$) and mush ($x>s(t)$), the free boundary condition is

$$E(s(t)+t,t)\big[\dot{s}(t)-v_0(s(t))\big]+K_S\big[u'_m(s(t))-u_x^s\big]=0, \tag{3.6}$$

where $K_S=K(0-)$ and $u_x^s=u_x(s(t)-,t)$.

(B) When $\dot{s}(t)\leqq v_0(s(t))$, the free boundary condition between solid and mush is

$$u_x^S=u'_m(s(t)), \tag{3.7}$$

and the condition

$$R(0-)+K_Su''_m(s(t))>0 \tag{3.8}$$

is needed.

Conditions (3.6), (3.7) can be summarized as

$$E(s(t)+,t)\big[\dot{s}(t)-v_0(s(t))\big]^++K_S\big[u'_m(s(t))-u_x^S\big]=0. \tag{3.9}$$

This unilateral form of the free boundary condition can be derived using the boundary point principle (see [6] for the details). Parallel results are found for an L-M interface, while on a L-S interface the usual Stefan condition is valid.

4. An example. In this section we produce a simple example in which, despite of the absence of volumetric heat sources, mushy regions appear as an effect of a variable melting temperature with nonvanishing second derivative. We note that in this definition of classical solution, possible effects of superheating are not taken into account.

Indeed our classical solution originates from a weak solution under suitable regularity requirements, and the definition of the "energy" function (2.1) is a constitutive law excluding nonequilibrium effects (see [6], [10], [11] and [13] for a discussion of this aspect).

We consider the slab $0<x<1$ assuming that no body sources are present ($r=0$), that the specific heat is constant throughout the slab (say, $c=1$) and that the conductivity has two constant values K_L and K_S in the liquid and in the solid, respectively. Therefore we define the function $K(E)$ as follows:

$$K(E)=\begin{cases} K_L, & E>\Lambda, \\ K_S+hE, & E\in(0,\Lambda), \quad h=\dfrac{K_L-K_S}{\Lambda}, \\ K_S, & E<0. \end{cases}$$

We will assume that the melting temperature is given by

$$u_m(x)=-1+x^2, \tag{4.1}$$

and that the initial and boundary conditions are

$$u(x,0)=-1+\frac{x^2}{2}, \qquad x\in(0,1), \tag{4.2}$$

$$u_x(0,t)=0, \qquad t>0, \tag{4.3}$$

$$u(1,t)=-\tfrac{1}{2} \qquad t>0. \tag{4.4}$$

We assume that this problem has a classical solution and we will derive some a priori information on the qualitative behaviour of the solution.

Our first result is aimed at characterizing the region S.

PROPOSITION 1. *There exists a function* $s(t)\geqq 0$, *such that* $S=\{(x,t)\colon s(t)<x<1, 0<t<T\}$. *Moreover,* $s(t)\not\equiv 0$ *in any neighborhood of* $t=0$.

The proof of this proposition is essentially based on the maximum principle. The next step will be to prove that a mushy region appears from the very beginning.

PROPOSITION 2. *There exists a function* $z(t)\geqq 0$, $z\not\equiv s$, *in any neighborhood of* $t=0$, $z(t)\leqq s(t)$, *such that* $u(x,t)=u_m(x)$ *in the set* $z(t)\leqq x\leqq s(t)$, $0<t<T$.

Proof. Suppose that there exist a $t_1>0$ and a region $R=\{y(t)<x<s(t), 0<t<t_1\}$ with $y(t)\geqq 0$, such that R is a component of $L\cap\{t<t_1\}$.

The maximum principle ensures that if $y(t)>0$, then the curve $x=y(t)$ is an interface between M and L. Following the discussion of the preceding section, it can be seen that the condition $u_m''>0$ implies $\dot{y}(t)<v_0(y(t))=-2hy(t)$, $t\in(0,t_1)$, thus contradicting $y(t)>0$, $y(0)=0$. Let us show that also assuming $y(t)$ identically equal to zero is contradictory. In such a case the pair (s,u) would be a solution of

$$\begin{aligned} &K_L u_{xx}-u_t=0, \quad 0<x<s(t), && 0<t<t_1, \\ &u_x(0,t)=0, && 0<t<t_1, \\ &u(s(t),t)=u_m(s(t)), && 0<t<t_1, \\ &u(x,t)>u_m(x), \quad 0<x<s(t), && 0<t<t_1. \end{aligned}$$

Using Green's theorem we find

$$\int_0^{s(t)} u(x,t)\,dx=\int_0^t K_L u_x(s(t')-,t')\,dt'+\int_0^t u(s(t'),t')\dot{s}(t')\,dt'. \tag{4.5}$$

The last integral is nothing but $\int_0^{s(t)} u_m(x)\,dx$, hence

$$0 \leqq \int_0^t K_L u_x^L dt' = -\Lambda s(t) + \int_0^t K_S u_x^s dt'$$

$$\leqq -\Lambda s(t) + \int_0^t K_S u'_m(s(t'))\,dt'.$$

This inequality is false, since $u'_m(s) = 2s$. The argument applies also in any interval $(t_a, t_b) \subset (0, T)$ in which $s(t) > 0$, $s(t_a) = 0$. Thus the only possibility is that the curve $x = s(t)$ is an interface between M and S.

By means of similar arguments it is possible to prove that the liquid region does not appear at $t = 0$. More precisely, we have

PROPOSITION 3. *There exists a time $t_0 > 0$ such that*

$$z(t) \equiv 0 \quad \text{in } (0, t_0).$$

Therefore, in this time interval $(0, t_0)$, we will have nothing but one free boundary separating regions M (for $x < s(t)$) and S (for $x > s(t)$).

Having concluded our a priori analysis, we can pass to the proof of our existence theorem. According to the results of §3, to find the interface between S and M together with the temperature in the region S we have to solve the following problem: find a $T > 0$, and $s(t) \in C[0, T] \cap C^1(0, T)$ and a $u(x,t) \in C^{2,1}(D_T) \cap C(\bar{D}_T)$ ($D_T \equiv \{(x,t)\colon s(t) < x < 1, 0 < t < T\}$) such that

$$\begin{aligned}
&K_S u_{xx} - u_t = 0 && \text{in } D_T,\\
&s(0) = 0,\\
&u(x,0) = -1 + \frac{x^2}{2}, && 0 < x < 1,\\
&u(1,t) = -\frac{1}{2}, && 0 < t < T,\\
&u(s(t),t) = -1 + s^2(t), && 0 < t < T,
\end{aligned} \tag{4.6}$$

and that

$$K_S\left[u_x(s(t),t) - 2s(t)\right] = E(s(t)-,t)\left[\dot{s}(t) - v_0(s(t))\right], \tag{4.7}$$

in the case $\dot{s}(t) < -2hs(t)$. In the case $\dot{s}(t) \geqq -2hs(t)$, instead of (4.7) the following condition has to be satisfied

$$u_x(s(t),t) = 2s(t). \tag{4.8}$$

If we disregard the condition $\dot{s} \geqq -2hs$, problem (4.6), (4.8) has a unique classical solution, such that $\lim_{t \to 0} \dot{s}(t) = +\infty$ and $s(t) \leqq 1 - (2)^{-1/2}$. This conclusion can be obtained using the results of [5]. Such solution is also a solution to the original problem if $h \geqq 0$ (the condition $\dot{s} \geqq -2hs$ is automatically satisfied). Otherwise a time $\bar{T} > 0$ exists such that $\dot{s}(\bar{T}) = -2hs(\bar{T})$, beyond which the solution has to be continued in a different way. The possibility of such a continuation will not be investigated here.

The determination of M requires first the solution of

$$E_t - 2hxE_x = 2(K_S + hE) \tag{4.9}$$

with the condition

$$E(s(t),t) = 0, \qquad 0 < t \quad \text{if } h > 0, \qquad 0 < t < \bar{T} \quad \text{if } h < 0 \tag{4.10}$$

and the constraint $E \in (0, \Lambda)$. This problem is easily solved and the curves $x = q_c(t)$ where $E = c$ are found to have positive and bounded slope for any value of c in $(0, \Lambda]$.

The L-M interface, $x = z(t)$, is expected to start from the point $(0, T_L)$, with $T_L = (2h)^{-1} \ln(1 + h\Lambda/K_S)$ for $h \neq 0$ and $T_L = \Lambda/(2K_S)$ for $h = 0$, and enter the region bounded by the curves $x = s(t)$ and $x = q_L(t)$.

On such an interface the energy balance condition is

$$K_L[u_x(z(t)-,t) - 2z(t)] = -[\Lambda - E(z(t)+,t)][\dot{z}(t) - v_0(z(t))], \qquad t > T_L. \tag{4.11}$$

As a matter of fact, it can be shown that the alternative condition $u_x(z(t)-,t) = 2z(t)$, valid when $\dot{z}(t) < v_0(z(t))$, never occurs

We have the following.

THEOREM. *The free boundary problem for the heat conduction equation in the liquid phase with free boundary conditions* (4.11), $z(T_L) = 0$ *and*

$$u(z(t),t) = -1 + z^2(t) \tag{4.12}$$

has one unique classical solution in some interval (T_L, T).

The proof of this theorem is complicated by the fact that the apparent latent heat in (4.11) vanishes for $t = T_L$. The main tool employed is to construct a sequence of monotone approximations by solving the following problems.

$$\begin{aligned}
&K_L u^{(i)}_{xx} - u^{(i)}_t = -2K_L \quad \text{in } D^{(i)}_c \equiv \{(x,t): 0 < x < z^{(i)}(t), T_L < t < T^{(i)}_c\}, \\
&z^{(i)}(T_L) = \frac{1}{i}, \\
&u^{(i)}(x, T_L) = 0, \qquad x \in \left(0, \frac{1}{i}\right), \\
&u^{(i)}_x(0,t) = 0, \qquad t \in (T_L, T^{(i)}_c), \\
&u^{(i)}(z^{(i)}(t),t) = 0, \qquad t \in (T_L, T^{(i)}_c), \\
&K_L u^{(i)}_x(z^{(i)}(t),t) = -L^*(z^{(i)}(t),t)[\dot{z}^{(i)}(t) + 2hz^{(i)}(t)], \qquad t \in (T_L, T^{(i)}_c).
\end{aligned} \tag{4.13}$$

In (4.13) the following symbols are used:

1) i is an integer not less than some i_0 such that $E(1/i_0, T_L)$ is defined and takes a value in $(0, \Lambda)$;

2) c is a constant in $(0, \Lambda)$ such that the level curve $x = q_c(t)$ hits $t = T_L$ on the left of $x = 1/i$;

3) $L^*(x,t) = \Lambda - E(x,t)$;

4) $T^{(i)}_c$ is the first instant (greater than T_L) such that the curve $x = z^{(i)}(t)$ hits either $x = q_c(t)$ or the boundary of the region where E is defined.

The convergence proof goes through the following steps.

First, one can prove as in [12] that for any $i > i_0$ there exists a constant $c_0(i)$ such that $T^{(i)}_c$ is uniformly estimated from below by some $T_0 > T_L$ for any $c \in (c_0, L)$. Hence the following definition makes sense

$$z(t) = \lim z^{(i)}(t), \qquad t \in (T_L, T_0).$$

Next, it can be shown that if the functions $q_\Lambda(t)$ and $z(t)$ do not coincide in any neighborhood of $t = T_L$, then $q_\Lambda(t) < z(t)$ in (T_L, T_0). The argument of the proof is based upon the construction of suitable barriers for $z^{(i)}$.

By means of a comparison technique it can then be proved that the convergence of the sequence $\{z^{(i)}\}$ is uniform in $[T_L, T_0]$ and that the limit function $z(t)$ is locally Lipschitz continuous. Moreover, let $\hat{u}(x,t)$ be the solution of the equation $K_L\hat{u}_{xx}-\hat{u}_t=-2K_L$ in the domain $0<x<z(t)$, $T_L<t<T_0$, subjected to the boundary conditions $\hat{u}_x(0,0)=0$, $\hat{u}(z(t),t)=0$. The maximum principle ensures that $\hat{u}(x,t)$ is the uniform limit of the sequence $\{u^{(i)}(x,t)\}$.

At this point it can be concluded that the pair (z,u), with $u=\hat{u}-1+x^2$, solves the free boundary problem in the liquid phase, i.e., that (4.11) is satisfied. To this end it is convenient to reformulate (4.11) in an integral form, using Green's identity, namely

$$\int_{T_L}^{t}\int_0^{z(t)} 2K_L\,dx\,dt' + A(z(t),t) + \int_{T_L}^{t} A_1(z(t),t)\,dt' - \int_0^{z(t)} \hat{u}(x,t)\,dx, \qquad t\in(T_L,T_0), \tag{4.14}$$

where

$$A(x,t) = -\int_0^x L^*(x',t)\,dx',$$

$$A_1(x,t) = -2hxL^*(x,t) + \int_0^x L_t^*(x',t)\,dx'.$$

In the present case (4.14) is seen to be equivalent to (4.11). Writing the corresponding equality for each of the pairs $(z^{(i)},u^{(i)})$ and passing to the limit shows the validity of (4.14).

As a consequence, the theorem will be proved if we can guarantee that the functions $q_\Lambda(t)$ and $z(t)$ are not identical in a right neighborhood of $t=T_L$. Assuming this is false, we immediately get a contradiction, since L^* and $\hat{u}_x$ vanish on $x=z(t)$, where $\hat{u}$ attains its minimum.

REFERENCES

[1] D. R. ATTHEY, *A finite difference scheme for melting problems*, J. Inst. Math. Appl., 13 (1974), pp. 353–366.

[2] A. BERMUDEZ AND C. SAGUEZ, *Etude numérique d'un problème de solidification d'un alliage*, INRIA report 137, Le Chesnay, 1982.

[3] C. M. BRAUNER, M. FREMOND AND B. NICOLAENKO, *A new homographic approximation to multiphase Stefan problems*, in [7], II, pp. 365–379.

[4] A. B. CROWLEY AND J. R. OCKENDON, *On the numerical solution of an alloy solidification problem*, Int. J. Heat Mass Transfer, 22 (1979), pp. 941–947.

[5] A. FASANO AND M. PRIMICERIO, *A critical case for the solvability of Stefan-like problems*, Math. Meth. Appl. Sci., 5 (1983), pp. 1–13.

[6] _____, *Mushy regions with variable temperature in melting processes*, to appear.

[7] A. FASANO AND M. PRIMICERIO, eds., *Free Boundary Problems: Theory and Applications*, Research Notes in Mathematics, 78, 79, Pitman, London, 1983.

[8] M. FREMOND, *Frost action in soils*, in [7], I, pp. 191–211.

[9] R. GORENFLO AND K. H. HOFFMANN, eds., *Applied Nonlinear Functional Analysis*, Lang, Frankfurt, 1982.

[10] A. A. LACEY AND M. SHILLOR, *The existence and stability of regions with super-heating in the classical two-phase one-dimensional Stefan problem with heat sources.*, IMA J. Appl. Math., 30 (1983), pp. 215–230.

[11] A. A. LACEY, J. R. OCKENDON AND A. B. TAYLER, *Modelling mushy regions*, to appear.

[12] M. PRIMICERIO, *Mushy regions in phase-change problems*, in [9], pp. 251–269.

[13] L. RUBINSTEIN, *On mathematical models for solid-liquid zones in a two-phase monocomponent system and in binary alloys*, in [7], I, pp. 275–282.

[14] D. G. WILSON, A. D. SOLOMON AND V. ALEXIADES, *A shortcoming of the explicit solution for the binary alloy solidification problem*, to appear.

SIAM J. MATH. ANAL.
Vol. 17, No. 1, January 1986

009

STANDING WAVE SOLUTIONS FOR A SYSTEM DERIVED FROM THE FITZHUGH–NAGUMO EQUATIONS FOR NERVE CONDUCTION*

GENE A. KLAASEN† AND ENZO MITIDIERI‡

Abstract. We focus our attention on the reaction-diffusion system $u_t = D_1\Delta u + f(u) - v$, $v_t = D_2\Delta v + \varepsilon(u - \gamma v)$ where $f(u) = u(1-u)(u-a)$, $0 < a < \frac{1}{2}$ and D_1, D_2, ε, γ are positive constants. The parameter γ is chosen large so that the associated dynamic equations ($D_1 = D_2 = 0$) have three constant solutions two of which are stable. The authors establish necessary and sufficient conditions that the Dirichlet problem for this system possesses two nontrivial time independent solutions.

1. Introduction.

We investigate the system

$$\begin{aligned} u_t &= D_1\Delta u + f(u) - v, \\ v_t &= D_2\Delta v + \varepsilon(u - \gamma v), \end{aligned} \tag{1.1}$$

where $\Delta \equiv \sum_{i=1}^{n} \partial^2/\partial x_i^2$, $n \geqq 1$, $t \geqq 0$, $(x_1, \cdots, x_n) \in \Omega \subseteq \mathbb{R}^n$, $f(u) = u(1-u)(u-a)$, $0 < a < \frac{1}{2}$, $D_1 > 0$, $D_2 > 0$, $\varepsilon > 0$, $\gamma > 0$. Equations (1.1) are an extension of the simpler FitzHugh–Nagumo [7], [13] equations, namely

$$\begin{aligned} u_t &= u_{xx} + f(u) - v, \\ v_t &= \varepsilon(u - \gamma v). \end{aligned} \tag{1.2}$$

The FitzHugh–Nagumo system serves as a prototype for nerve conduction and other chemical and biological systems. The interested reader is referred to [8], [15] for a review of results obtained to this date.

The space independent system (i.e., the dynamics of (1.1) and (1.2)) consists of the equations

$$\begin{aligned} u_t &= f(u) - v, \\ v_t &= \varepsilon(u - \gamma v). \end{aligned} \tag{1.3}$$

Of particular interest to us is the case of large γ for which (1.3) have three steady state solutions, two of which are stable and one unstable (see Fig. 1). This "bistability" phenomenon is present in a number of chemical and biological models [6], [13], [21].

Recently, Terman and Rinzel [18] have made a detailed mathematical and numerical study of the case γ large in a simplification of (1.2),

$$\begin{aligned} u_t &= u_{xx} + H(u - a) - u - v, \\ v_t &= \varepsilon(u - \gamma v), \end{aligned} \tag{1.4}$$

*Received by the editors August 2, 1983, and in revised form June 25, 1984.

†Department of Mathematics, University of Tennessee, Knoxville, Tennessee 37996. Present address: Department of Mathematics, Calvin College, Grand Rapids, Michigan 49506. The work of this author was supported in part by the National Science Foundation under grant MCS 8002948 and by the Italian CNR.

‡Istituto Di Matematica, Universitá Degli Studi Di Trieste, 34100 Trieste, Italy. The work of this author was supported by the Italian CNR.

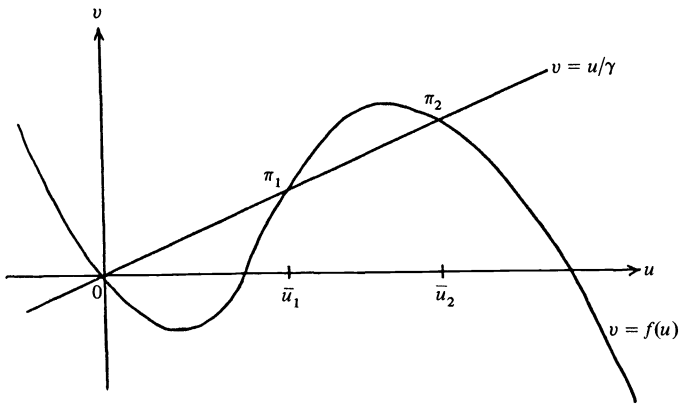


FIG. 1

where

$$H(u-a)=\begin{cases}1 & \text{if } u>a,\\ 0 & \text{if } u<a,\end{cases} \qquad -\infty<u<\infty.$$

For appropriate values of the parameters they find wave fronts, wave backs and pulse solutions of (1.4). One important result of their analysis is that $\varepsilon\gamma^2<1$ is a necessary and sufficient condition for the existence of a pulse solution.

In this paper we investigate the more complicated system (1.1). The addition of diffusion terms to the v equation reflects the possibility that more than one species may diffuse. For example, Tuckwell and Miura [19] have proposed a model for spreading cortical depression in the brain. In their system there are three steady state solutions and more than one species diffuses. Also, Koga and Kuramoto [10] examine a two diffusion modification of (1.4), namely

$$\begin{aligned} u_t &= D_1u_{xx}+H(u-a)-u-v,\\ v_t &= D_2v_{xx}+bu-cv, \qquad -\infty<x<\infty. \end{aligned} \tag{1.5}$$

They show that if the inhibiting diffusion D_2 is sufficiently large, (1.5) have 2 nonconstant steady state solutions, one of which is linearly stable. They also argue that such standing waves are likely in bistable excitable systems in which the inhibiting diffusion is large enough to halt the formation of a traveling wave.

Rothe and de Mottoni [16], [17] consider a system somewhat similar to (1.1) but with $D_2=\varepsilon=\gamma=1$ and a function f which satisfies $f(-u)=-f(u)$ and $f'(0)>0$. In contrast to our system, their model has a single constant solution which exhibits the diffusion driven instability property which Turing suggests as the dynamics of pattern formation. Moreover, since our f is not odd we cannot apply the Ljusternik–Schnirelmann theory as they do to obtain existence of steady state solutions.

Standing waves of (1.1) have been found when $n=1$. Ermentrout and Hastings [5] show when $\gamma>0$ is small that there are two standing waves. Klaasen and Troy [9] argue for $\gamma>0$ large and $\frac{1}{2}-a>0$ small, $n=1$ that system (1.1) has a standing wave solution and infinitely many periodic solutions.

Our main goal is to show the existence of standing wave solutions for system (1.1) when $n>1$. Our space domains are finite balls with sufficiently large radius, i.e., $B_R(0)=\{(x_1,\cdots,x_n)\,|\,\Sigma_{i=1}^n x_i^2\leqq R^2\}$, where $R>0$ is large.

In §2 we state our main results. In §3 we give the proofs of our theorems.

2. Statement of main results. We investigate (1.1) for the existence of steady state solutions on bounded domains Ω in $\mathbb{R}^N$ with zero boundary values. Such solutions $(u(x), v(x))$ solve the system

$$\begin{aligned} -D_1\Delta u &= f(u)-v, \\ -D_2\Delta v &= \varepsilon u-\varepsilon\gamma v, \end{aligned} \tag{2.1}$$

on Ω with $u=0$ and $v=0$ on $\partial\Omega$. By rescaling the equations using the transformations $x\to x/\sqrt{D_1}$, $\delta D_2=\varepsilon D_1$, we obtain the equivalent system

$$-\Delta u = f(u)-v \quad \text{on } \Omega, \qquad u=0, \quad v=0 \quad \text{on } \partial\Omega. \tag{2.2}$$

$$-\Delta v = \delta u-\delta\gamma v \tag{2.3}$$

We assume that $0<a<\frac{1}{2}$ and require $\gamma>4/(1-a)^2$. This latter condition guarantees that (2.2) have three constant solutions. The first of these, the so-called rest state, is given by $\pi_0=(0,0)$. The other two are denoted by $\pi_i=(\bar{u}_i,\bar{v}_i)$, $i=1,2$, where $0<\bar{u}_1<\bar{u}_2$ and $0<\bar{v}_1<\bar{v}_2$. By solving the equation $f(u)=u/\gamma$ it is easy to see (see Fig. 1) that

$$\bar{u}_i=\frac{1+a}{2}+\frac{(-1)^i}{2}\sqrt{(1+a)^2-4\left(a+\frac{1}{\gamma}\right)}, \qquad i=1,2.$$

If we assume $\partial\Omega$ is of class $C^{2+\alpha}$, where $\alpha\in(0,1)$, then, following [11], [16], [17], the system (2.2), (2.3) can be uncoupled since the boundary value problem

$$-\Delta v+\delta\gamma v=\delta u, \qquad v=0 \quad \text{on } \partial\Omega, \tag{2.4}$$

defines a transformation $v=B(u)$, where B can be viewed as a bounded invertible linear transformation from $C^{\alpha}(\bar{\Omega})$ into $C^{2+\alpha}(\bar{\Omega})$ or from $L^2(\Omega)$ into $H_0^1(\Omega)$ in the case of weak solutions. Substituting $v=B(u)$ into (2.2) we obtain an equivalent single operator equation

$$-\Delta u+B(u)=f(u) \quad \text{on } \Omega, \qquad u=0 \quad \text{on } \partial\Omega. \tag{2.5}$$

Finally if we define

$$h(u)=(u-a)(u-1),$$

then (2.5) becomes

$$-\Delta u+B(u)=-uh(u) \quad \text{on } \Omega, \qquad u=0 \quad \text{on } \partial\Omega. \tag{2.6}$$

Thus u, v is a solution for the Dirichlet problem (2.2), (2.3) if and only if $v=B(u)$ and u is a solution of (2.6).

The following lemma asserts that classical solutions of (2.6) are a priori bounded.

LEMMA 1. *Let $b_0>0$ be chosen so that $h(u)>1/\gamma$ whenever $|u|>b_0$. Then every C^2-solution u of* (2.6) *satisfies*

$$|u(x)|\leqq b_0, \qquad |Bu(x)|\leqq\frac{b_0}{\gamma} \quad \text{on } \Omega.$$

The variational approach to solving (2.6) is enhanced by the a priori bounds for classical solutions as established in Lemma 1. We use these bounds to modify h as follows. Let $g \in C'(\mathbb{R})$ be defined with the restrictions

$$\text{(2.7)} \quad \begin{aligned} &\text{(i)} \quad g(u)=h(u), && |u| \leqq b_0, \\ &\text{(ii)} \quad g(u)>\frac{1}{\gamma}, && |u|>b_0, \\ &\text{(iii)} \quad g \text{ and } g' \text{ are bounded on } \mathbb{R}. \end{aligned}$$

Then $-ug(u)$ is a modification of $f(u)$ and if $F(u)=\int_0^u f(s)\,ds$ then the function $\tilde{F}(u)=\int_0^u -sg(s)\,ds$ is a modification of F. Let the functional Φ be defined on $H_0^1(\Omega)$ by

$$\text{(2.8)} \qquad \Phi(u)=\frac{1}{2}\int_\Omega |\nabla u|^2 dx+\frac{1}{2}\int_\Omega uB(u)\,dx-\int_\Omega \tilde{F}(u)\,dx.$$

By standard critical point theory, see [10], critical points of Φ are weak solutions of the Dirichlet problem

$$\text{(2.9)} \qquad -\Delta u+B(u)=-ug(u) \quad \text{on } \Omega \qquad u=0 \quad \text{on } \partial\Omega.$$

Ambrosetti and Rabinowitz [2] and Ambrosetti [1] prove that if there are constants c_1, c_2 and p such that

$$|k(s)| \leqq c_1+c_2|s|^p$$

where

$$\text{(2.10)} \qquad \begin{aligned} &p>1 && \text{if } n=2, \quad \text{but} \\ &1 \leqq p<\frac{n+2}{n-2} && \text{if } n>2, \end{aligned}$$

then weak solutions of Dirichlet problems for $-\Delta u=k(u)$ on domains $\Omega \subset \mathbb{R}^n$ with $\partial\Omega \in C^{2+\alpha}$, $0<\alpha<1$, are of class $C^{2+\alpha}$. Our choice of g is such that there are constants c_1 and c_2 with $|sg(s)| \leqq c_1+c_2|s|$ and hence condition (2.10) follows for $k(u)=-ug(u)$. Because of the smoothing effect of B and (2.10), a standard "boot strap" argument (see [1] or [2]) can be applied to conclude that if

$$\text{(2.11)} \qquad \Omega \subset \mathbb{R}^n, \qquad \partial\Omega \in C^{2+\alpha} \quad \text{for some } 0<\alpha<1,$$

then weak solutions of (2.9) are classical solutions. But then Lemma 1 and (2.7) imply that these solutions are solutions of (2.6). Throughout the remainder of our discussion we will assume our domains Ω satisfy (2.11).

THEOREM 1. *Suppose* $0<a<\frac{1}{2}$, $\gamma>9/(2a^2-5a+2)$. *There exists an* $R_0>0$ *such that if* Ω *contains a ball of radius* R_0 *then the Dirichlet problem* (2.2), (2.3) *has a nontrivial solution pair* $u_1, v_1=B(u_1)$ *of class* $C^{2+\alpha}(\Omega)$ *which satisfies*

$$\inf_{H_0^1(\Omega)} \Phi(u)=\Phi(u_1)<0.$$

THEOREM 2. *If the hypotheses of Theorem* 1 *hold then a second nontrivial pair of solutions* $u_2, v_2=B(u_2)$ *of the Dirichlet problem* (2.2), (2.3) *exists of class* $C^{2+\alpha}(\Omega)$ *and satisfies*

$$\inf_{\sigma\in\Sigma} \max_{0\leqq t\leqq 1} \Phi(\sigma(t))=\Phi(u_2)>0,$$

where $\Sigma=\{\sigma \in C([0,1]; H_0^1(\Omega)) \mid \sigma(0)=0, \sigma(1)=u_1\}$.

The hypotheses of Theorem 1 and Theorem 2 are to a certain degree necessary as the following theorem indicates.

THEOREM 3. *The Dirichlet problem* (2.2), (2.3) *fails to have a weak nontrivial solution on* $\Omega = B_R(0)$ *if any one of the following hypotheses is assumed*:

(i) δ, γ *are fixed positive numbers and* $R > 0$ *is sufficiently small*;

(ii) $\delta\gamma^2 \geqq 1$, $\gamma < 4/(1-a)^2$ *and any* $R > 0$;

(iii) $\delta\gamma^2 < 1$, $2\sqrt{\delta} - \delta\gamma > (1-a)^2/4$ *and any* $R > 0$.

The hypothesis

$$\gamma > \frac{9}{2a^2 - 5a + 2} \tag{2.12}$$

present in Theorems 1 and 2 is quite natural. Firstly, since

$$\frac{9}{2a^2-5a+2} > \frac{4}{(1-a)^2},$$

condition (2.12) implies that the dynamics of (2.2) are bistable as mentioned earlier. Secondly, condition (2.12) is equivalent to the inequality

$$\int_0^{\bar{u}_2} \left(f(s) - \frac{s}{\gamma} \right) ds > 0, \tag{2.13}$$

where $\bar{u}_2$ is the first coordinate of the third constant solution $\pi_2 = (\bar{u}_2, \bar{v}_2)$ of (2.2). Inequality (2.13) is known to be a necessary condition for the existence of nontrivial solutions of Dirichlet problems associated with the single equation $-\Delta u = f(u) - u/\gamma$; see Berestycki and Lions [3].

In part (iii) of Theorem 3 the assumption that $\gamma < 4/(1-a)^2$ is implicit in the inequality $2\sqrt{\delta} - \delta\gamma > (1-a)^2/4$ since it is always true that $1/\gamma \geqq 2\sqrt{\delta} - \delta\gamma$. Hence a necessary condition for the existence of nontrivial solutions of the Dirichlet problem (2.2), (2.3) is that $\gamma > 4/(1-a)^2$.

3. Proofs. This section contains the proofs of Lemma 1 and Theorems 1, 2 and 3 of §2.

Proof of Lemma 1. This proof is essentially due to Lazer and McKenna [11]. Let u be a C^2-solution of (2.5) with Dirichlet boundary conditions. Let $\max_\Omega |u(x)| = b_1$. By applying the maximum principle to $-\Delta v = -\delta\gamma u + \delta u$ we see that $|Bu(x)| = |v(x)| \leqq b_1/\gamma$ on Ω. Suppose for contradiction that $b_1 > b_0$. If there exists $x_1 \in \Omega$ such that $u(x_1) = b_1$, then $\Delta u(x_1) \leqq 0$ and

$$u(x_1) < \gamma h(u(x_1))u(x_1) = \gamma\left[\Delta u(x_1) - B(u(x_1))\right] \leqq -\gamma B(u(x_1)) \leqq u(x_1),$$

which is impossible. Similarly there is no $x_2 \in \Omega$ such that $u(x_2) = -b_1$ and hence $b_1 > b_0$ is false and the lemma is proved.

Before proving Theorem 1 we introduce some notation and establish the validity of several useful lemmas.

On the Hilbert space $L^2(\Omega)$ the inner product is denoted by

$$(u, v) = \int_\Omega uv\, dx$$

and the norm is denoted by

$$\|u\|_2 = (u, u)^{1/2}.$$

If the Hilbert space is $H_0^1(\Omega)$, then the inner product will be

$$((u,v))=\int_\Omega(\nabla u\cdot\nabla v)\,dx$$

and the norm

$$\|u\|=((u,u))^{1/2}.$$

We seek nontrivial solutions of

$$-\Delta u+B(u)=-uh(u) \quad \text{on } \Omega, \qquad u=0 \quad \text{on } \partial\Omega. \tag{3.1}$$

Recalling the discussion following Lemma 1 of §2, we will assume $\Omega\subseteq R^n$ satisfies $\partial\Omega\in C^{2+\alpha}$, where $0<\alpha<1$, and obtain as a consequence that weak solutions u of (3.1) are of class $C^{2+\alpha}(\Omega)$ and $u,v=B(u)$ is a pair of classical solutions of (2.2), (2.3). Also weak solutions of (3.1) are critical points of the functional

$$\Phi(u)=\frac{1}{2}\int_\Omega|\nabla u|^2+\frac{1}{2}\int_\Omega uB(u)-\int_\Omega\tilde{F}(u), \tag{3.2}$$

where $\tilde{F}(u)=-\int_0^u sg(s)\,ds$.

LEMMA 2. *Let B be defined by* (2.4). *Then*

$$\|B\|\leqq\frac{1}{\gamma}, \quad \textit{where } \|B\|=\inf_{u\neq0}\frac{\|Bu\|_2}{\|u\|_2}.$$

Proof. If $u\in L^2(\Omega)$ and $v=B(u)$ then $(-\Delta v)v+\delta\gamma v^2=\delta uv$ and hence $\int_\Omega|\nabla v|^2+\delta\gamma\int_\Omega v^2=\delta\int uv\leqq\delta\|u\|_2\|v\|_2$. Consequently, $\delta\gamma\|Bu\|_2^2=\delta\gamma\int_\Omega v^2\leqq\delta\|u\|_2\|Bu\|_2$ and the lemma easily follows. □

Let $B_R=\{x\in\mathbb{R}^n\,|\,\|x\|<R\}$, $C_R=B_R-B_{R-1}$ and for any $\Omega\subseteq\mathbb{R}^n$ let $|\Omega|=\int_\Omega dx$.

LEMMA 3. *If $R>0$ is sufficiently large then there exists a $u\in H_0^1(B_R)$ such that $\Phi(u)<0$.*

Proof. For $R>0$ define $u_R\in H_0^1(B_R)$ by

$$u_R=\begin{cases}\bar{u}_2, & 0\leqq\|x\|\leqq R-1,\\ (R-\|x\|)\bar{u}_2, & R-1\leqq\|x\|\leqq R,\end{cases}$$

where $(\bar{u}_2,\bar{v}_2)$ is the third constant solution of (2.2). Then

$$\begin{aligned}\Phi(u_R)&=\frac{1}{2}\int_{B_R}|\nabla u_R|^2+\frac{1}{2}\int_{B_R}u_RB(u_R)-\int_{B_R}\tilde{F}(u_R)\\ &=\frac{1}{2}\int_{C_R}|\nabla u_R|^2+\frac{1}{2}\int_{B_R}u_RB(u_R)-\int_{B_{R-1}}\tilde{F}(\bar{u}_2)-\int_{C_R}\tilde{F}(u_R)\\ &\leqq\frac{\bar{u}_2^2}{2}|C_R|+\frac{1}{2\gamma}\int_{B_R}u_R^2-\tilde{F}(\bar{u}_2)|B_{R-1}|+c_1|C_R|,\end{aligned}$$

where c_1 is a positive constant independent of R. Recall that $|B_R|=k_nR^n$, where k_n is a constant depending only on the dimension n of the space. Thus there is a constant l_n such that $|C_R|\leqq l_nR^{n-1}$. Continuing the inequality we have that there exists a constant K_n such that

$$\Phi(u_R)\leqq K_nR^{n-1}+k_n\left(\frac{\bar{u}_2^2}{2\gamma}-\tilde{F}(\bar{u}_2)\right)R^n.$$

Since $a<\bar{u}_2<1<b_0$, $+\tilde{F}(\bar{u}_2)>\bar{u}_2^2/2$ if $\gamma>9/(2a^2-5a+2)$ and hence for R sufficiently large $\Phi(u_R)<0$. See (2.13).

Proof of Theorem 1. First we wish to argue that Φ, defined by (3.2), is bounded below on $H_0^1(\Omega)$. From (2.7) we have that $g(s)>1/\gamma$ for $|s|>b_0$ and hence there exists an $M>0$ such that $\tilde{F}(u)=\int_0^u sg(s)\,ds>-M$ for all $u\in\mathbb{R}$. Consequently $\int_\Omega \tilde{F}(u)\,dx>-M|\Omega|$ on $H_0^1(\Omega)$. If the second equation of (2.2) is multiplied by $v=B(u)$ and integrated we obtain $\int_\Omega|\nabla v|^2+\delta\gamma\int_\Omega|v|^2=\delta\int_\Omega uB(u)$ and hence $\int_\Omega uB(u)\geqq 0$ for all $u\in H_0^1(\Omega)$. Consequently $\Phi(u)\geqq -M|\Omega|$ on $H_0^1(\Omega)$.

Secondly, since $\int_\Omega\tilde{F}(u)\,dx>-M|\Omega|$, $\phi(u)\geqq\frac{1}{2}\|u\|^2-M|\Omega|$ on $H_0^1(\Omega)$ and hence $\Phi(u)\to\infty$ as $\|u\|\to\infty$.

Finally, by standard variational arguments Φ is weakly lower semicontinuous and consequently Φ attains a global minimum on $H_0^1(\Omega)$ at some critical point u_1 which is a weak solution of (3.1); see Vainberg [19] or Ambrosetti [1]. Lemma 3 implies that $\Phi(u_1)<0$ and u_1 is nontrivial. By the standard "bootstrap" arguments referred to earlier, $u_1\in C^{2+\alpha}(\Omega)$ and is a classical solution.

To prove Theorem 2 we use the following version of the mountain pass theorem of Ambrosetti and Rabinowitz [2]; see Mawhin [12].

THEOREM. *Let $\Phi\in C^1(E,\mathbb{R})$, where E is a Hilbert space and assume there exists $u_0\in E$, $u_1\in E$, $r>0$, $k>0$ satisfying the following conditions*:

(i) $\|u_0-u_1\|>r$;

(ii) $\Phi(u_0)<k$, $\phi(u_1)<k$;

(iii) $\Phi(u)\geqq k$ *for* $\|u-u_0\|=r$.

Let $\Sigma=\{\sigma\in C([0,1],E)\mid\sigma(0)=u_0,\ \sigma(1)=u_1\}$ *and define*

$$c=\inf_{\sigma\in\Sigma}\max_{u\in\sigma([0,1])}\Phi(u).$$

If Φ satisfies the Palais–Smale condition, then c is a critical value for Φ.

In the proof of Theorem 2, $E=H_0^1(\Omega)$, $u_0=0$; u_1 is described in Theorem 1. Since B is a compact operator the argument of Ambrosetti and Rabinowitz [2] or Ambrosetti [1] will show that the function Φ defined in (3.2) satisfies the Palais–Smale condition. The remainder of the hypotheses of the mountain pass theorem will be satisfied when we prove the following lemma.

LEMMA 4. *There exist $r>0$, $\rho>0$ such that $\Phi(u)>0$ for all $0<\|u\|\leqq r$ and $\phi(u)\geqq\rho$ for all $\|u\|=r$.*

Proof. Define $k(u)=ug(u)+au$ so that (2.9) becomes

$$-\Delta u+B(u)=-au+k(u). \tag{3.3}$$

Let $K(u)=\int_0^u k(s)\,ds$. Then from (2.6) and (2.7) we conclude that $k(u)=(1+a)u^2-u^3$ for $|u|\leqq b_0$ and since $|sg(s)|\leqq c_1+c_2|s|$ for $s\in\mathbb{R}$ we see that k satisfies (2.10). From this property Ambrosetti and Rabinowitz [2] show that $\int_\Omega K(u)\,dx=0(\|u\|^2)$ at $u=0$. Hence

$$\begin{aligned}\Phi(u)&=\frac{1}{2}\int_\Omega|\nabla u|^2+\int_\Omega B(u)u-\int_\Omega\tilde{F}(u)\\&\geqq\frac{1}{2}\int_\Omega|\nabla u|^2+\frac{a}{2}\int_\Omega u^2-\int_\Omega K(u)\\&\geqq\varepsilon\|u\|^2+O\big(\|u\|^2\big)\quad\text{for some fixed }\varepsilon>0\end{aligned}$$

and the lemma follows. □

In Theorem 3 we prove the nonexistence of nontrivial solutions to the boundary value problem

$$\begin{aligned} -\Delta u + B(u) &= f(u) && \text{on } B_0(R), \\ u &= 0 && \text{on } \partial B_0(R). \end{aligned} \tag{3.4}$$

Several lemmas are required to establish the validity of Theorem 3.

Consider the eigenvalue problem:

$$\begin{aligned} -\Delta u + Bu &= \mu u && \text{on } B_0(R), \\ u &= 0 && \text{on } \partial B_0(R). \end{aligned} \tag{3.5}$$

LEMMA 5. *If μ is an eigenvalue of* (3.5), *then $\mu = \lambda + \delta/(\lambda + \delta\gamma)$, where λ is an eigenvalue of*

$$\begin{aligned} -\Delta u &= \lambda u && \text{on } B_0(R), \\ u &= 0 && \text{on } \partial B_0(R). \end{aligned} \tag{3.6}$$

Proof. Suppose (μ, u) is an eigenvalue eigenfunction pair for equation (3.5) on $L^2(\Omega)$. Since $B = \delta(\delta\gamma - \Delta)^{-1}$ we operate on both sides of (3.5) by $(\delta\gamma - \Delta)$ to obtain

$$\begin{aligned} &\left((-\Delta)^{-2} + \delta\gamma(-\Delta) + \delta\right)u = (\mu\delta\gamma - \mu\Delta)u \quad \text{or} \\ &\left[(-\Delta)^2 + (\delta\gamma - \mu)(-\Delta) + (\delta - \mu\delta\gamma)\right]u = 0. \end{aligned} \tag{3.7}$$

By factoring the quadratic operator in $-\Delta$ we obtain

$$(-\Delta + a)(-\Delta + b)u = 0 \tag{3.8}$$

for appropriate complex numbers a and b. Both a and b must be real numbers for otherwise they are complex conjugates and both $(-\Delta + a)$ and $(-\Delta + b)$ are invertible. This would imply that $u = 0$ contradicting the fact that u is in eigenfunction. Hence a and b are both real and (3.8) implies that one is an eigenvalue of $-\Delta$ with eigenfunction u. If we suppose a is an eigenvalue then from (3.7) we conclude that $a^2 + \delta\gamma a + \delta - \mu\delta\gamma = 0$ and upon solving for μ we obtain the required relationship $\mu = a + \delta/(a + \delta\gamma)$.

LEMMA 6. *Let $\nu(\lambda) \equiv \lambda + \delta/(\lambda + \delta\gamma)$. Then on $(-\delta\gamma, \infty)$, $\nu(\lambda)$ has a unique positive minimum at $\lambda_0 = \sqrt{\delta} - \delta\gamma$ with $\nu(\lambda_0) = 2\sqrt{\delta} - \delta\gamma$. Moreover ν is increasing on $\left[\sqrt{\delta} - \delta\gamma, \infty\right)$ and $\lim_{\lambda \to \infty} \nu(\lambda) = \infty$. This lemma is obvious.*

LEMMA 7. *Let μ_1 be the first eigenvalue of* (3.5). *If the* BVP (3.1) *has a nontrivial solution, then $\mu_1 \leqq \frac{1}{4}(1-a)^2$.*

Proof. Let $u \in L^2(B_R(0))$ be a nontrivial solution of BVP (3.1), and let $C \equiv -\Delta + B$. Then by the Rayleigh–Ritz representation of the first eigenvalue of (3.5), see [4], we have $\mu_1 = \inf(Cv, v)/\|v\|_2^2$, where infimum is taken over all $v \in L^2(B_R(0))$, $v \neq 0$. Since $u \in L^2(B_R(0))$, we have

$$\begin{aligned} \mu_1 \|u\|_2^2 \leqq (Cu, u) = (f(u), u) &= \int_{B_R(0)} \left[-u^4 + (1+a)u^3 - au^2\right] \\ &= \int_{B_R(0)} u^2\left[-u^2 + (1+a)u - a\right] \leqq \int_{B_R(0)} u^2\left(\frac{1-a}{2}\right)^2 \\ &= \frac{(1-a)^2}{4}\|u\|_2^2 \end{aligned}$$

or $\mu_1 \leqq \frac{1}{4}(1-a)^2$.

Proof of Theorem 3. First we prove part (i).

Let $0<a<\frac{1}{2}$, δ, γ, be fixed. If $\lambda_1(R)$ is the first eigenvalue of (3.6), then $\lim_{R\downarrow 0}\lambda_1(R)=\infty$; see [4]. Hence, from Lemma 6, there exists $R_0>0$ such that if $0<R<R_0$ then $\nu(\lambda_1(R))>\frac{1}{4}(1-a)^2$ and $\lambda_1(R)>\lambda_0$. Hence the first eigenvalue $\mu_1(R)$ of (3.5) satisfies

$$\mu_1(R)=\lambda_n+\frac{\delta}{\lambda_n+\delta\gamma}=\nu(\lambda_n)\geqq\nu(\lambda_1(R))$$

and by Lemma 7 no nontrivial solution of (3.1) exists.

To prove part (ii), since $\delta\gamma^2\geqq 1$ and $1/\gamma>\frac{1}{4}(1-a)^2$, then in Lemma 6, $\lambda_0\leqq 0$ and hence $\mu(\lambda)$ is increasing on $[0,\infty)$. Thus

$$\frac{1}{\gamma}=\mu(0)<\mu(\lambda)$$

for all $\lambda>0$ and hence for some n, $\mu_1(R)=\mu(\lambda_n)>\mu(0)=1/\gamma>\frac{1}{4}(1-a)^2$ and by Lemma 7 no nontrivial solution of (3.1) exists.

Finally part (iii) follows by an argument similar to part (ii). For if u is a nontrivial solution of (3.1) for some $R>0$ then from Lemma 7 we have

$$\mu_1(R)\leqq\frac{1}{4}(1-a)^2<2\sqrt{\delta}-\delta\gamma=\min_{\lambda\geqq 0}\mu(\lambda)$$

which contradicts Lemma 5. Hence no nontrivial solution of (3.1) exists. □

REFERENCES

[1] A. AMBROSETTI, *Topics in critical point theory*, lecture notes given at Autumn Course on Variational Methods in Analysis and Mathematical Physics, International Center for Theoretical Physics, Trieste, Italy, 1981.

[2] A. AMBROSETTI AND P. H. RABINOWITZ, *Dual variational methods in critical point theory and applications*, J. Funct. Anal. 14 (1973), pp. 349–381.

[3] H. BERESTYCKI AND P. L. LIONS, *Une methode locale pour l'existence de solutions positives de problèmes semi-linéaires elliptiques dans R^n*, J. d'Analyse Math., 38 (1980), pp. 144–187.

[4] R. COURANT AND D. HILBERT, *Methods of Mathematical Physics*, Vol. 2, Interscience, New York, 1962.

[5] G. B. ERMENTROUT AND S. P. HASTINGS, private communication.

[6] P. G. FIFE, *Pattern formation in reacting and diffusing systems*, J. Chem. Phys., 64 (1976), pp. 554–564.

[7] R. FITZHUGH, *Impulses and physiological states in theoretical models of nerve membrane*, Biophys. J., 1 (1961), pp. 445–466.

[8] S. P. HASTINGS, *Some mathematical problems from neurobiology*, Amer. Math. Monthly, 82 (1975), pp. 881–895.

[9] G. KLAASEN AND W. TROY, *Standing wave solutions of a system of reaction-diffusion equations derived from the FitzHugh-Nagumo equations*, SIAM J. Appl. Math., 44 (1984), pp. 96–110.

[10] S. KOGA AND Y. KURAMOTO, *Localized patterns in reaction-diffusion systems*, Prog. Theoret. Phys., 63 (1980), pp. 106–121.

[11] A. LAZER AND P. MCKENNA, *On steady state solutions of a system of reaction-diffusion equations from biology*, Nonlinear Anal., 6 (1982), pp. 523–530.

[12] J. MAWHIN, *Variational methods and boundary value problems for ordinary differential equations*, Lecture Notes at Colorado State University, July, 1982.

[13] J. NAGUMO, S. YOSHIZAWA AND S. ARIMOTO, *Bistable transmission lines*, IEEE Trans. Circuit Theory, CT-12 (1965), pp. 400–412.

[14] P. ORTOLEVA AND J. ROSS, *Theory of propagation of discontinuities in kinetic systems with multiple time scales: fronts, front multiplicity and pulses*, J. Chem. Phys., 63 (1975), pp. 3398–3408.

[15] J. RINZEL, *Impulse propagation in excitable systems*, in Dynamics and Modelling of Reactive Systems, W. E. Stewart, W. H. Ray and C. C. Conley, eds., Academic Press, New York, 1980, pp. 259–291.

[16] R. ROTHE AND P. DE MOTTONI, *A simple system of reaction-diffusion equations describing morphogenesis* I. *Asymptotic behavior*, Ann. Mat. Pura Appl., 122 (1979), pp. 141–157.

[17] F. ROTHE, *Global existence of branches of stationary solutions for a system of reaction-diffusion equations from biology*, Nonlinear Anal., 5 (1981), pp. 487–598.

[18] D. TERMAN AND J. RINZEL, *Propagation phenomena in a bistable reaction diffusion system*, SIAM J. Appl. Math., 42 (1982), pp. 1111–1137.

[19] H. TUCKWELL AND R. MIURA, *A mathematical model for cortical depression*, Biophys. J., 23 (1978), pp. 257–276.

[20] M. VAINBERG, *Variational Methods in the Study of Nonlinear Operators*, Holden-Day, San Francisco, 1964.

[21] H. R. WILSON AND J. D. COWAN, *Excitatory and inhibitory interactions in localized populations of model neurons*, Biophys. J., 12 (1972), pp. 1–24.

SIAM J. MATH. ANAL.
Vol. 17, Nol. 1, January 1986

010

REGULARIZATION FOR A CLASS OF NONLINEAR EVOLUTION EQUATIONS*

JAMES F. EPPERSON†

Abstract. We establish estimates of the regularization error for a class of nonlinear degenerate parabolic evolution equations, which includes the Stefan problem in enthalpy form.

1. Introduction. In this paper we consider the regularization of nonlinear parabolic equations in the form

$$u_t = \Delta f(u) + F, \tag{1}$$

where $f'(\xi) \geqq 0$, and equality holds for some $\xi \in \text{Range}(u)$. Our ultimate goal is to establish estimates for the error due to changing from f to f_ε, $f_\varepsilon'(\xi) \geqq \varepsilon > 0$ for all ξ.

Particular equations of the form (1) have been studied by many authors, including Friedman [8], Cannon and Hill [5], Brezis [3], Jerome [9] and Alexiades and Cannon [2], all of whom established existence and uniqueness in appropriate Sobolev spaces.

The question of regularization error for equations of the form (1) has been most closely tied to the Stefan problem. In [7] we established an $O(\varepsilon^{1/4})$ estimate for the L^2 error in temperature (corresponding to $f(u)$ in (1)), where ε is the regularizing parameter. This was extended in [6] to $O(\varepsilon^{1/2})$. Jerome and Rose [10] had earlier obtained a similar $O(\varepsilon^{1/2})$ result for a different regularization.

The importance of the regularization error lies in its effect on the error in numerical approximation. If the regularization f_ε is such that $f_\varepsilon'(\xi) \geqq \varepsilon > 0$, then typical numerical error estimates ([6], [10]) are proportional to positive powers of ε^{-1}. There thus exists an ε^* for which the approximation and regularization errors are balanced. The smaller the regularization error, then, the smaller the total error.

Our main result is Theorem 4.1, which characterizes the regularization error in both u and $f(u)$ for a fairly broad class of regularizations. This is followed by Theorem 4.2, which gives conditions under which the regularization error for the Stefan problem is $O(\varepsilon)$.

To fix notation, let $\Omega \subset \mathbb{R}^n$ be an open bounded domain with Lipschitz boundary $\partial\Omega = \Gamma$. For $T > 0$, $Q_T = \Omega \times (0, T]$ defines a space-time cylinder. The spaces $H^r(D)$ denote the usual Sobolev spaces of functions over D, with the standard norms $\|\cdot\|_{r,D}$ [1]. The notation $L^p(X)$ for $L^p(0, T; X)$ will be used for brevity, with the norm being written $\|\cdot\|_{L^p(X)}$. Finally, the letter C will be used to denote generic positive constants. Only where distinction is important will any attempt be made to distinguish between different constants.

2. The nondegenerate case. Consider the evolution problem

$$u_t = \Delta f(u) + F \quad \text{in } \Omega, \quad t > 0, \tag{2}$$

$$u = 0 \quad \text{on } \Gamma, \quad t > 0, \tag{3}$$

$$u(0) = u_0 \quad \text{on } \Omega, \quad t = 0, \tag{4}$$

*Received by the editors October 25, 1983, and in revised form May 10, 1984.
† Department of Mathematics, University of Georgia, Athens, Georgia 30602.

where we assume

(5) $f \in C(\mathbb{R})$ and is piecewise C^1,

(6) $f(0)=0, \qquad 0<\sigma \leqq f'(\xi) \leqq \kappa < \infty \quad$ all ξ.

We remark that the choice of homogeneous boundary data is not a restriction, since for smooth enough boundary data an appropriate translation of f and modification of F can always be done to achieve $u=0$ on Γ. We also take $F \in L^2(Q_T)$ and $u_0 \in H^1(\Omega)$, although the case $u_0 \in L^2(\Omega)$ will also be considered.

It follows from (5)–(6) that, for any $\xi, \eta \in \mathbb{R}^1$,

(7) $$(f(\xi)-f(\eta))\cdot(\xi-n) \geqq C\sigma(\xi-\eta)^2$$

and

(8) $$(f(\xi)-f(\eta))\cdot(\xi-\eta) \geqq C(f(\xi)-f(\eta))^2,$$

where C is independent of σ.

We begin by establishing certain a priori estimates for smooth solutions to (2)–(4). By "smooth" we mean a function in the space

$$V=\left\{ \varphi \in L^2(0,T;\, H^2(\Omega)\cap H_0^1(\Omega)) \,\middle|\, \varphi_t \in L^2(Q_T) \right\}.$$

LEMMA 2.1. *Let $\varphi \in V$ solve* (2)–(4) *a.e. in Q_T. Then*;

(i) $\|\varphi\|_{L^\infty(L^2)} \leqq CM_1$;

(ii) $\|\varphi\|_{L^2(H_0^1)} \leqq C\sigma^{-1/2}M_1$;

(iii) $\|f(\varphi)\|_{L^2(H_0^1)} \leqq CM_1$;

(iv) $\|(f(\varphi))_t\|_{L^2(L^2)}+\|f(\varphi)\|_{L^\infty(H_0^1)} \leqq CM_2$;

(v) $\|\varphi_t\|_{L^2(L^2)} \leqq C\sigma^{-1/2}M_2$;

(vi) $\|f(\varphi)\|_{L^2(H^2)} \leqq C\sigma^{-1/2}M_2$,

where

$$M_1=\|u_0\|_0+T^{1/2}\|F\|_{L^2(L^2)}, \qquad M_2=\|\nabla f(u_0)\|_0+\|F\|_{L^2(L^2)}.$$

The constants in (i)–(vi) *are all independent of σ.*

Proof. We use standard energy arguments. Multiply the equation by φ, integrate over space and time, and use the Poincaré inequality to get

(9) $$\|\varphi(t)\|_0^2+\sigma\int_0^t\|\varphi(\tau)\|_1^2 d\tau \leqq C\left(\|u_0\|_0^2+\int_0^t\int_\Omega|\varphi F|dx\,d\tau\right);$$

alternately, we can use (6) to obtain

(10) $$\|\varphi(t)\|_0^2+\int_0^t\|f(\varphi(\tau))\|_1^2 d\tau \leqq C\left(\|u_0\|_0^2+\int_0^t\int_\Omega|\varphi F|dx\,d\tau\right).$$

The last term on the right can be bounded by the Hölder inequality applied twice:

$$\begin{aligned}\int_0^t\int_\Omega|\varphi F|dx\,d\tau &\leqq \int_0^t\|\varphi(\tau)\|_0\|F(\tau)\|_0 d\tau\\ &\leqq \|\varphi\|_{L^\infty(L^2)}\int_0^t\|F(\tau)\|_0 d\tau\\ &\leqq t^{1/2}\|\varphi\|_{L^\infty(L^2)}\|F\|_{L^2(L^2)}.\end{aligned}$$

Results (i) and (ii) follow immediately from this last bound applied to (9); result (iii) follows when it is applied to (10).

If we now multiply (2) by $(f(\varphi))_t$ and again integrate twice, we get

$$\int_0^t\int_\Omega f'(\varphi)(\varphi_t)^2 dx\,d\tau+\int_0^t\int_\Omega \nabla(f(\varphi))_t\nabla f(\varphi)\,dx\,d\tau=\int_0^t\int_\Omega (f(\varphi))_t F\,dx\,d\tau,$$

which becomes

$$(11)\quad \int_0^t \|f(\varphi(\tau))_t\|_0^2\,d\tau+\|\nabla f(\varphi(t))\|_0^2\leqq C\left(\|\nabla f(u_0)\|_0^2+\int_0^t\int_\Omega |f(\varphi(\tau))_t F(\tau)|dx\,d\tau\right),$$

or, alternately

$$(12)\quad \sigma\int_0^t \|\varphi_t(\tau)\|_0^2\,d\tau+\|\nabla f(\varphi(t))\|_0^2\leqq C\left(\|\nabla f(u_0)\|_0^2+\int_0^t\int_\Omega |f(\varphi(\tau))_t F(\tau)|dx\,d\tau\right).$$

Part (iv) follows immediately from (11); part (v) then follows from (12) with (iv) used to bound the last integral on the right. Finally, part (vi) is established by multiplying (2) by $-\Delta f(\varphi)$ and bounding the φ_t term using (v).

THEOREM 2.2. *Suppose* $u_0\in H^1(\Omega)$ *and* $F\in L^2(Q_T)$. *Then there exists a unique* $u\in H^1(Q_T)$, *with* $f(u)\in L^2(H^2(\Omega))$, *solving* (2)–(4).

Proof. Existence and uniqueness were established in [5] for weak solutions in the sense that

$$(13)\qquad \int_0^T\int_\Omega \left[u\varphi_t+f(u)\Delta\varphi\right]dx\,dt+\int_\Omega u_0\varphi(0)\,dx+\int_0^T\int_\Omega F\varphi\,dx\,dt=0$$

for all $\varphi\in V$, $\varphi(T)=0$ a.e. in Ω. This analysis assumed homogeneous Neumann data and a specific choice of nonlinearity f, but can be easily generalized to our case. Further, the estimates of Lemma 2.1 imply the extended regularity $u\in H^1(Q_T)$, $f(u)\in L^2(0,T;H^2(\Omega))$, i.e., u is a strong solution in that (2) holds a.e. in Q_T.

THEOREM 2.3. *Suppose only that* $u_0\in L^2(\Omega)$ *and* $F\in L^2(Q_T)$. *Then there exists* u, *a unique weak solution* (*in the sense of* (13)) *to* (2)–(4), *such that* $u\in L^\infty(L^2(\Omega))$ *and* $f(u)\in L^2(H_0^1(\Omega))$.

Proof. The theorem follows from Theorem 2.2 and Lemma 2.1(iii) by taking a sequence $\{v_0^n\}$ in $H^1(\Omega)$ converging in $L^2(\Omega)$ to u_0.

3. The degenerate case. We now consider the problem (2)–(4), with hypothesis (6) changed to

$$(6')\qquad f(0)=0,\qquad 0\leqq f'(\xi)\leqq\kappa<\infty,\quad \text{all } \xi\in\mathbb{R}^1.$$

Because $\sigma=0$, many of the bounds from Lemma 2.1 no longer apply. Thus, the whole existence—uniqueness question becomes much more delicate. Direct proofs of existence of weak solutions are possible ([3], [8], for example), but we prefer here to establish existence as the limiting case, as $\sigma\to 0$, of (2)–(4), (6).

Consider the sequence of problems:

$$(14)\qquad u_t^{(n)}=\Delta\left(f(u^{(n)})+\sigma_n u^{(n)}\right)+F \quad \text{in } \Omega,\quad t>0,$$

$$(15)\qquad u^{(n)}=0 \quad \text{on } \Gamma,\quad t>0,$$

$$(16)\qquad u^{(n)}(0)=u_0 \quad \text{in } \Omega.$$

Each $u^{(n)}$ is the solution of a problem of the form (2)–(4), (6) and, if

$$\lim_{n\to\infty}\sigma_n=0,$$

then we can in fact construct the solution to the degenerate problem.

THEOREM 3.1. *Let* $F\in L^2(Q_T)$ *and* $u_0\in H^1(\Omega)$. *Then there exists a unique weak solution of* (2)–(4), (6′); *moreover*,

$$u\in L^\infty\big(0,T;\, L^2(\Omega)\big),$$
$$f(u)\in L^\infty\big(0,T;\, H_0^1(\Omega)\big),$$
$$f(u)_t\in L^2\big(0,T;\, L^2(Q_T)\big).$$

Proof. Consider the sequences $\{u^{(n)}\}$, $\{f_n(u^{(n)})\}$, defined by (14)–(16), where

$$f_n(\xi)=f(\xi)+\sigma_n\xi.$$

By Theorem 2.2, then, the sequences are well-defined and

$$\|u^{(n)}\|_{L^2(Q_T)}\leqq C,\qquad \big\|f_n(u^{(n)})\big\|_{H^1(Q_T)}\leqq C,$$

where C is independent of n for $\lim_{n\to\infty}\sigma_n=0$. Hence there exist subsequences such that

$$u^{(m)}\to\bar u \quad \text{weakly in } L^2(Q_T), \tag{17}$$

$$f_m(u^{(m)})\to\bar f \quad \text{strongly in } L^2(Q_T). \tag{18}$$

Clearly, if $\bar f=f(\bar u)$, then we have established existence, for the same analysis as in [5] will apply. Since f is continuous and monotone, it suffices to show that

$$\big(\bar u-v,\bar f-f(v)\big)\geqq 0,\quad \text{all } v\in L^2(Q_T). \tag{19}$$

So, we consider

$$\big(\bar u-v,\bar f-f(v)\big)=\big(\bar u-u^{(m)},\bar f-f(v)\big)+\big(u^{(m)}-v,\bar f-f_m(u^{(m)})\big)$$
$$+\big(u^{(m)}-v,f(u^{(m)})-f(v)\big)+\sigma_m\big(u^{(m)}-v,u^{(m)}\big).$$

In the limit, the first two terms go to zero by (17) and (18); the third term is positive since f is monotone; and the last goes to zero since σ_m does. Thus (19) holds, and so $\bar f=f(\bar u)$. Existence then follows.

For uniqueness, let $v\in L^2(Q_T)$ be a second solution. Then

$$\int_0^T\int_\Omega\big[(u-v)\varphi_t+\big(f(u)-f(v)\big)\Delta\varphi\big]\,dx\,dt=0$$

for all $\varphi\in V$. Factor out the $(u-v)$ to get

$$\int_0^T\int_\Omega(u-v)[\varphi_t+e\Delta\varphi]\,dx\,dt=0,\quad \text{where } e=\frac{f(u)-f(v)}{u-v}\geqq 0$$

and $e\in L^\infty(Q_T)$. Consider then the final value problem

$$\varphi_t^{(\varepsilon)}+(e+\varepsilon)\varphi^{(\varepsilon)}=\psi \quad \text{in } \Omega,\quad t<T,$$
$$\varphi^{(\varepsilon)}=0 \quad \text{on } \Gamma,\quad t<T,$$
$$\varphi^{(\varepsilon)}(T)=0, \quad \text{in } \Omega,$$

for $\psi\in C_0^\infty(Q_T)$, $\varepsilon>0$. By [11, pp. 179–180], this has a unique solution $\varphi^{(\varepsilon)}\in V$. Thus, for any $\psi\in C_0^\infty(Q_T)$,

$$\int_0^T\int_\Omega(u-v)\psi\,dx\,dt=\int_0^T\int_\Omega(u-v)\varepsilon\Delta\varphi^{(\varepsilon)}\,dx\,dt.$$

Again from [11, pp. 179–180], we have that

$$\|\varepsilon^{1/2}\Delta\varphi^{(\varepsilon)}\|_{L^2(L^2)} \leqq C$$

for C independent of ε. Hence, we have

$$\left|\int_0^T\int_\Omega (u-v)\psi\,dx\,dt\right| \leqq C\varepsilon^{1/2}\|u-v\|_{L^2(Q_T)}$$

for any $\psi \in C_0^\infty(Q_T)$, any $\varepsilon > 0$, which implies $u = v$ a.e. in Q_T.

As before, a solution for $u_0 \in L^2(\Omega)$ is possible.

THEOREM 3.2. *Let $F \in L^2(Q_T)$ and $u_0 \in L^2(\Omega)$. Then there exists a unique solution u to* (2)–(4), (6′); *moreover*

$$u \in L^\infty\big(L^2(\Omega)\big), \qquad f(u) \in L^2\big(H_0^1(\Omega)\big).$$

Proof. Again, take a sequence of $\{v_0^{(j)}\}$, $v_0^{(j)} \in H^1(\Omega)$, $v_0^{(j)} \to u_0$ in $L^2(\Omega)$.

4. Regularization. Consider the degenerate problem (2)–(4), (6′) and the associated regularization (14)–(16). What bounds, if any, can be placed on the errors $u - u^{(n)}$ and/or $f(u) - f_n(u^{(n)})$? In connection with numerical work for the Stefan problem, this question has been addressed before ([6], [7], [10]). The best estimates so far are ([6], [10]) $O(\varepsilon^{1/2})$ in $\|f(u) - f_\varepsilon(u^\varepsilon)\|_{L^2}$, where ε is the regularizing parameter ($f'(\xi) \geqq \varepsilon$). Here we generalize and improve upon some of these results.

For a given f satisfying (5) and (6′), define f_ε by

$$f_\varepsilon(\xi) = f(\xi) + e_\varepsilon(\xi),$$

where e_ε is continuous and such that

$$e_\varepsilon(0) = 0, \qquad f_\varepsilon'(\xi) \geqq \varepsilon \quad \text{for } \varepsilon < \varepsilon_0.$$

Let u^ε solve the corresponding problem (2)–(4). Then

THEOREM 5.1. *If $F \in L^2(Q_T)$ and $u_0 \in L^2(\Omega)$, then*

(i) $$\|u - u^\varepsilon\|_{L^2(L^2)} \leqq \varepsilon^{-1}\|e_\varepsilon(u)\|_{L^2(L^2)};$$

(ii) $$\|f(u) - f_\varepsilon(u^\varepsilon)\|_{L^2(L^2)} \leqq \big(1 + (\kappa/\varepsilon)^{1/2}\big)\|e_\varepsilon(u)\|_{L^2(L^2)}.$$

Proof. Let $E: H^{-1}(\Omega) \to H_0^1(\Omega)$ be the solution operator for the boundary value problem

$$-\Delta\varphi = \psi \quad \text{in } \Omega, \qquad \varphi = 0 \quad \text{on } \Gamma.$$

Then the PDE's can be written as

$$Eu_t^\varepsilon + f_\varepsilon(u^\varepsilon) = EF, \qquad u^\varepsilon(0) = u_0,$$
$$Eu_t + f(u) = EF, \qquad u(0) = u_0.$$

Subtract, multiply by $(u - u^\varepsilon)$, and integrate to get

$$\frac{1}{2}\|E^{1/2}(u - u^\varepsilon)\|_0^2(T) + \int_0^T\int_\Omega \big(u - u^\varepsilon, f(u) - f_\varepsilon(u^\varepsilon)\big)\,dx\,dt = 0,$$

hence

$$\int_0^T\int_\Omega \big(u - u^\varepsilon, f_\varepsilon(u) - f_\varepsilon(u^\varepsilon)\big)\,dx\,dt \leqq \int_0^T\int_\Omega e_\varepsilon(u)(u - u^\varepsilon)\,dx\,dt.$$

Thus, from (7), and the Schwarz inequality,

$$\varepsilon\|u-u^\varepsilon\|^2_{L^2(L^2)} \leqq \|e_\varepsilon(u)\|_{L^2(L^2)}\|u-u^\varepsilon\|_{L^2(L^2)},$$

which proves (i). Then, using (8),

$$\|f_\varepsilon(u)-f_\varepsilon(u^\varepsilon)\|^2_{L^2(L^2)} \leqq \kappa\|e_\varepsilon(u)\|_{L^2(L^2)}\|u-u^\varepsilon\|_{L^2(L^2)},$$

so

$$\|f_\varepsilon(u)-f_\varepsilon(u^\varepsilon)\|_{L^2(L^2)} \leqq \left(\frac{\kappa}{\varepsilon}\right)^{1/2}\|e(u)\|_{L^2(L^2)},$$

hence (ii) follows, using the triangle inequality.

The most obvious regularization is the global choice, $e_\varepsilon(u)=\varepsilon u$, for all u. In this case, Theorem 5.1 yields

$$\|f(u)-f_\varepsilon(u^\varepsilon)\|_{L^2(L^2)} = O(\varepsilon^{1/2})$$

for all ε, but only that $\|u-u^\varepsilon\|_{L^2(L^2)}$ is bounded.

On the other hand, for certain cases a regularization e_ε is possible which provides higher order convergence.

THEOREM 5.2. *Suppose that the following conditions hold for ε sufficiently small*:

(a) $\|e_\varepsilon(u)\|_{L^\infty(Q_T)} \leqq C\varepsilon$;

(b) $\mathrm{vol}(\mathrm{supp}\, e_\varepsilon(u)) \leqq C\varepsilon$.

Then, for ε sufficiently small:

(i) $\|u-u^\varepsilon\|_{L^2(L^2)} \leqq C\varepsilon^{1/2}$;

(ii) $\|f(u)-f_\varepsilon(u^\varepsilon)\|_{L^2(L^2)} \leqq C\varepsilon$.

Proof. Conditions (a) and (b) directly imply

$$\|e_\varepsilon(u)\|_{L^2(L^2)} \leqq C\varepsilon^{3/2},$$

from which both (i) and (ii) follow, using Theorem 5.1.

Consider, then, the Stefan problem in enthalpy form, in which case f has the general form

$$f(\xi)=\begin{cases}\xi+1, & 1\leqq\xi,\\ 0, & 0<\xi<1,\\ \xi, & \xi\leqq 0.\end{cases}$$

Define $e_\varepsilon(\xi)$ by

$$e_\varepsilon(\xi)=\begin{cases}0, & 1+\sqrt{\varepsilon}\leqq\xi,\\ \varepsilon-\sqrt{\varepsilon}(\xi-1), & 1\leqq\xi\leqq 1+\sqrt{\varepsilon},\\ \varepsilon\xi, & 0\leqq\xi\leqq 1,\\ 0, & \xi\leqq 0.\end{cases}$$

Clearly, Theorem 5.1 applies, with $\varepsilon_0=1$; moreover, condition (a) of Theorem 5.2 also holds.

Now suppose that the given problem also has a solution in the classical sense; for the Stefan problem this is the case under fairly mild and reasonable conditions on the data functions (see [4] and references therein). In this case, the degenerate region corresponds to the solid-liquid interface and hence has measure zero (it is an n-dimensional surface in $(n+1)$-space). Thus

$$\operatorname{vol}\{(x,t)|0 \leqq u(x,t) \leqq 1\} = 0.$$

But the degenerate region is very nearly the same as $\operatorname{supp} e_\varepsilon(u)$, and, in fact, as $\varepsilon \to 0$,

$$\operatorname{supp} e_\varepsilon(u) \to \{(x,t)|0 \leqq u \leqq 1\}.$$

Thus the hypothesis (b) of Theorem 5.2 becomes quite reasonable. This discussion then indicates that, for the classical Stefan problem, a regularization error that is $O(\varepsilon)$ is plausible although not yet rigorously established.

REFERENCES

[1] R. A. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.

[2] V. ALEXIADES AND J. R. CANNON, *Free boundary problems in solidification of alloys*, this Journal, 11 (1980), pp. 254–264.

[3] HAIM BREZIS, *On some degenerate nonlinear parabolic equations*, in Nonlinear Functional Analysis, Proc. Symposia in Pure Mathematics, 18, F. Browder, ed., American Mathematical Society, Providence, RI, 1970, pp. 28–38.

[4] J. R. CANNON, *Multiphase parabolic free boundary problems*, in Moving Boundary Problems, D. G. Wilson et al., eds., Academic Press, New York, 1978, pp. 3–24.

[5] J. R. CANNON AND C. D. HILL, *On the movement of a chemical interface*, Indiana Univ. Math. J., 20 (1970), pp. 429–454.

[6] J. F. EPPERSON, *Finite element methods for a class of nonlinear evolution equations*, SIAM J. Numer. Anal., 21 (1984), pp. 1066–1079.

[7] ———, *An error estimate for changing the Stefan problem*, SIAM J. Numer. Anal., 19 (1982), pp. 114–120.

[8] A. FRIEDMAN, *The Stefan problem in several space variables*, Trans. Amer. Math. Soc., 132 (1968), pp. 51–87.

[9] J. W. JEROME, *Nonlinear equations of evolution and a generalized Stefan problem*, J. Differential Equations 26 (1977), pp. 240–261.

[10] J. W. JEROME AND M. E. ROSE, *Error estimates for the multidimensional two phase Stefan problem*, Math. Comp., 39 (1982), pp. 377–414.

[11] O. A. LADYŽENSKAJA, V. A. SOLONNIKOV AND N. N. URAL'CEVA, *Linear and Quasilinear Equations of Parabolic Type*, American Mathematical Society, Providence, RI, 1968.

SIAM J. MATH. ANAL.
Vol. 17, No. 1, January 1986

011

BIFURCATION IN DOUBLY-DIFFUSIVE SYSTEMS I. EQUILIBRIUM SOLUTIONS*

WAYNE NAGATA[†‡] AND JAMES W. THOMAS[†]

Abstract. The problem of the steady flow arising when a layer of fluid with a dissolved solute is heated from below is considered. The problem is placed in a functional analytic setting where bifurcation of convective solutions is proved and the stability of these bifurcating branches is studied.

Key words. doubly-diffusive systems, bifurcation, stability

1. Introduction. This paper studies steady convective flow that can arise when a layer of fluid with a dissolved solute is heated from below. The effect of a stabilizing solute gradient can be overcome by introducing a destabilizing temperature gradient, resulting in convective motions. Under some conditions these motions can correspond to steady cellular flow, as in the case of pure thermal (Bénard) convection. However, it is also possible for the convective motions to be periodic in time.

Double-diffusive convection was studied in 1965 by Veronis [19], who simplified a model describing two-dimensional cellular flow (rolls) into a system of coupled nonlinear ordinary differential equations. Numerical solutions of these equations showed both equilibrium and periodic solutions. Later, Huppert and Moore [7] numerically integrated the full partial differential equations describing rolls, and found equilibrium, periodic and aperiodic solutions. More recently, in 1981, Da Costa, Knobloch and Weiss [6] solved the system of ordinary differential equations of Veronis numerically and discovered further behavior-successive period doubling and apparently chaotic solutions.

The initial appearance of convective motion as the destabilizing temperature gradient is increased is suggested by a study of the linearized stability of the motionless conduction solution which is globally stable for small temperature gradients. If the thermal diffusivity of the fluid is greater than its solute diffusivity, we have the diagram shown in Fig. 1 for the linearized stability of the conduction solution in two-dimensional parameter space [7, p. 826]. The parameter r is proportional to the destabilizing temperature gradient and the parameter s is proportional to the stabilizing solute concentration gradient. In the region below the curve ACD, the spectrum of the linearization about the conduction solution lies in the negative complex half-plane, and hence the conduction solution is asymptotically stable with respect to small perturbations. In the region above the curve ACD, the linearization has part of its spectrum in the positive complex half-plane, and hence the conduction solution is unstable. From bifurcation theory, we expect to find nontrivial solutions corresponding to convective flows near the linearized stability boundary ACD: (i) on the open line segment AC, the linearization about the conduction solution has a zero eigenvalue; (ii) on the open line segment CD, the linearization about the conduction solution has two conjugate pure imaginary eigenvalues; (iii) at the point C, the linearization has a degenerate zero eigenvalue. Case (i) is associated with the appearance of steady convective motion, case

*Received by the editors May 10, 1983, and in revised form June 11, 1984.
[†] Department of Mathematics, Colorado State University, Fort Collins, Colorado 80523.
[‡] Present address: Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912.

(ii) is associated with the appearance of convective motion in which the flow reverses itself periodically in time, and case (iii) is associated with both types of convection interacting with each other. When the thermal diffusivity is less than or equal to the solute diffusivity, then the line CD in Fig. 1 is absent and the linearized stability boundary of the conduction solution consists of the line AB. In this case only steady convection can bifurcate from the conduction solutions.

The bifurcation of nontrivial equilibrium solutions corresponding to steady convection is similar to that which occurs in the Bénard problem which models a layer of fluid heated from below with no solute present. The existence and stability of bifurcating equilibrium solutions has been shown for the Bénard problem by several authors, for example [8], [9]. A review of bifurcations in fluid flow, including the Bénard problem, was given by Kirchgässner in [13].

In this paper we treat case (i). We first describe the mathematical model, which was also the basis of the numerical studies mentioned above. Then in §3 we formulate the problem in a function space setting and apply a theorem of Crandall and Rabinowitz [4] to prove the existence of bifurcating equilibrium solutions corresponding to two- and three-dimensional cellular convection. In §4 we compute the linearized stability of the bifurcating solutions. We find that the three-dimensional solutions can be stable under the same conditions which would make two-dimensional roll-like solutions unstable.

In later papers—Parts II (this issue, pp. 114–127) and III—we will treat case (ii), involving the bifurcation of periodic solutions, and case (iii), involving the interaction between bifurcating equilibrium and periodic solutions.

2. Formulation of the problem. In this section we describe how to obtain the system of partial differential equations which we will use and formulate the boundary conditions.

The double-diffusive convection equations. Consider an infinite horizontal layer of an incompressible fluid of uniform height h where the upper surface is maintained at a constant temperature T_1 and solute concentration S_1, while the lower surface is maintained at constant temperature T_0 and solute concentration S_0. We assume that $T_0 > T_1$ and $S_0 > S_1$. The equations governing the motion of the fluid are taken to be [19]

$$
\begin{aligned}
&\frac{\partial}{\partial t}\mathbf{U} + (\mathbf{U}\cdot\nabla)\mathbf{U} = -\frac{1}{\rho_0}(\nabla P + \rho g \mathbf{e}) + \nu\Delta\mathbf{U},\\
&\frac{\partial}{\partial t}T + (\mathbf{U}\cdot\nabla)T = \kappa_T \Delta T,\\
&\frac{\partial}{\partial t}S + (\mathbf{U}\cdot\nabla)S = \kappa_S \Delta S,\\
&\nabla\cdot\mathbf{U} = 0,
\end{aligned}
\tag{2.1}
$$

for $\mathbf{x} = (x, y, z)$ in $\mathbb{R}^2 \times (0, h)$ where

$\nabla = (\partial/\partial x, \partial/\partial y, \partial/\partial z)$ is the gradient operator, $\Delta = \partial^2/\partial x^2 + \partial^2/\partial y^2 + \partial^2/\partial z^2$ is the Laplacian operator,

$\mathbf{U} = (U, V, W)$ is the fluid velocity at $\mathbf{x}$,

T is the fluid temperature at $\mathbf{x}$,

S is the solute concentration at $\mathbf{x}$,

ρ is the fluid density at **x**,
ρ_0 is the fluid density at the lower surface $z=0$, assume to be constant,
P is the fluid pressure at **x**,
g is the acceleration due to gravity,
e is the unit vector $(0,0,1)$,
ν is the kinematic viscosity,
κ_T is the thermal diffusivity,
κ_S is the solute diffusivity.

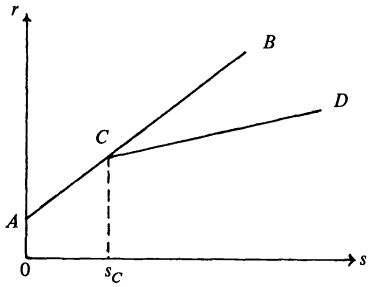


FIG. 1. *Linearized stability diagram for the conduction solution.*

We assume that the effects of the temperature and the solute appear only in the buoyancy force term $(\rho/\rho_0)g\mathbf{e}$, and that the fluid obeys an Oberbeck–Boussinesq equation of state

$$\rho=\rho_0[1-a(T-T_0)+b(S-S_0)], \tag{2.2}$$

where

$a=-\frac{1}{\rho}\frac{\partial\rho}{\partial T}$ is the thermal coefficient of volume expansion, assumed constant,

$b=\frac{1}{\rho}\frac{\partial\rho}{\partial S}$ is the solute coefficient of volume expansion, assumed constant.

We take ν, κ_T, κ_S to be constants, and we have the constant-gradient solution

$$\begin{aligned}
&\mathbf{U}^0=\mathbf{0},\\
&T^0=T_0-\left(\frac{T_0-T_1}{h}\right)z,\\
&S^0=S_0-\left(\frac{S_0-S_1}{h}\right)z,\\
&P^0=P_0-g\rho_0\left[z+\frac{a}{2}\left(\frac{T_0-T_1}{h}\right)z^2-\frac{b}{2}\left(\frac{S_0-S_1}{h}\right)z^2\right],
\end{aligned} \tag{2.3}$$

where P_0 is a constant. The solution (2.3) corresponds to pure conduction of heat and solute. We consider deviations from the conduction solution $\mathbf{u}=\mathbf{U}-\mathbf{U}^0$, $\theta=T-T^0$, $\zeta=S-S^0$, $p=P-P^0$. We also transform to dimensionless variables by making the

rescalings $\mathbf{x} \to h\mathbf{x}$, $t \to (h^2/\kappa_T)t$, $\mathbf{u} \to (\kappa_T/h)\mathbf{u}$, $\theta \to (T_0 - T_1)\theta$, $\zeta \to (S_0 - S_1)\zeta$, $p \to ((\rho_0 \nu \kappa_T)/h^2)p$. This leads to the dimensionless equations

$$
\begin{aligned}
&\frac{\partial}{\partial t}\mathbf{u} = \sigma(\Delta\mathbf{u} - \nabla p) + \sigma(r\theta - s\zeta)\mathbf{e} - (\mathbf{u}\cdot\nabla)\mathbf{u},\\
&\frac{\partial}{\partial t}\theta = \Delta\theta + w - (\mathbf{u}\cdot\nabla)\theta,\\
&\frac{\partial}{\partial t}\zeta = \tau\Delta\zeta + w - (\mathbf{u}\cdot\nabla)\zeta,\\
&\nabla\cdot\mathbf{u} = 0,
\end{aligned}
\tag{2.4}
$$

for $\mathbf{x} = (x, y, z)$ in the domain $\Omega_1 = \mathbb{R}^2 \times (0, 1)$, where $\mathbf{u} = (u, v, w)$ and

$$
r = \frac{ag(T_0 - T_1)h^3}{\kappa_T \nu}, \quad s = \frac{bg(S_0 - S_1)h^3}{\kappa_T \nu}, \quad \sigma = \frac{\nu}{\kappa_T}, \quad \tau = \frac{\kappa_S}{\kappa_T}.
$$

In the following sections we consider the system (2.4) with positive parameters r, s, σ and τ.

Boundary conditions. The upper and lower surfaces are maintained at constant temperatures and solute concentrations, so we must have

$$
\theta|_{z=0,1} = 0, \qquad \zeta|_{z=0,1} = 0 \tag{2.5}
$$

on the boundary of the domain. We assume that the fluid satisfies a stress-free condition on the boundary surfaces

$$
(\mathbf{u}\cdot\mathbf{N})|_{z=0,1} = 0, \qquad \left[\left(\nabla\mathbf{u} + (\nabla\mathbf{u})^T\right)\cdot\mathbf{N}\times\mathbf{N}\right]\Big|_{z=0,1} = 0,
$$

where $\mathbf{N}$ is the unit outward normal at a point on the boundary surface $(\nabla\mathbf{u})_{ij} = \partial u_i/\partial x_j$, and $(\cdot)^T$ denotes the matrix transpose. For the domain Ω_1 these conditions reduce to

$$
w|_{z=0,1} = 0, \quad \left.\frac{\partial u}{\partial z}\right|_{z=0,1} = 0, \quad \left.\frac{\partial v}{\partial z}\right|_{z=0,1} = 0. \tag{2.6}
$$

We note that the incompressibility condition $\nabla\cdot\mathbf{u} = 0$ and (2.6) together imply

$$
w|_{z=0,1} = \left.\frac{\partial^2 w}{\partial z^2}\right|_{z=0,1} = \left.\frac{\partial^4 w}{\partial z^4}\right|_{z=0,1} = \cdots = 0. \tag{2.7}
$$

Although these boundary conditions are difficult to approximate experimentally, they have the advantage that the eigenvalue problem for the linearization can be solved exactly.

Equations (2.4)–(2.6) comprise the problem which we study to find bifurcating nontrivial solutions. We observe that $\mathbf{u} = \mathbf{0}$, $\theta = 0$, $\zeta = 0$, $p =$ constant is always a solution of (2.4)–(2.6) for all positive values of the parameters r, s, σ and τ.

3. Bifurcation of equilibrium solutions. We begin this section by stating a theorem of Crandall and Rabinowitz which we will use to prove the existence of bifurcating equilibrium solutions of (2.4)–(2.6). We then examine the eigenvalue problem of the linearization to determine critical values of the parameters at which we expect to find bifurcations and set up the problem so that the critical (zero) eigenvalue is simple.

Finally, we define suitable function spaces and operators corresponding to the problem (2.4)–(2.6) and apply the bifurcation theorem.

The bifurcation theorem. Let us write (2.4)–(2.6) as a parameter-dependent evolution equation

$$\frac{du}{dt} = F(r,u) \tag{3.1}$$

in a suitable space of functions where s, σ and τ are considered fixed and the dependence of (3.1) on these parameters is suppressed. If $u=(\mathbf{u},\theta,\zeta)$, then $u\equiv 0$ is an equilibrium solution of (3.1) for all $r>0$. The following theorem gives conditions under which there exists a branch of nontrivial equilibrium solutions bifurcating from the trivial solution $u\equiv 0$; see [5, Lemma 1.1], proved in [4]. If L is a linear operator, we denote its null space by $N(L)$ and its range by $R(L)$.

THEOREM 1. *Let X and Y be real Banach spaces and let I be an open interval in $\mathbb{R}$. Suppose F: $I\times X\to Y$ is a C^{m+1} mapping with*

i) $F(r,0)=0$ *for all* $r\in I$,

ii) $\dim N(F_u(r_0,0)) = \operatorname{codim} R(F_u(r_0,0)) = 1$ *for some* $r_0\in I$,

iii) $F_{ru}(r_0,0)u_0 \notin R(F_u(r_0,0))$, *where* $N(F_u(r_0,0)) = \operatorname{span}\{u_0\}$.

Let Z be any complement of $\operatorname{span}\{u_0\}$ *in X. Then there exist an open interval I containing 0 and C^m functions r: $I\to\mathbb{R}$ and z: $I\to Z$ such that $r(0)=r_0$, $z(0)=0$ and if $u(\varepsilon)=\varepsilon u_0+\varepsilon z(\varepsilon)$, then $F(r(\varepsilon),u(\varepsilon))=0$ for all $\varepsilon\in I$. Moreover, the only nontrivial solutions of $F(r,u)=0$ near $(r_0,0)$ are of the form $(r(\varepsilon),u(\varepsilon))$ for some $\varepsilon\in I$.*

The eigenvalue problem for the linearization. To apply Theorem 1, we need to find function spaces X and Y such that zero is a simple eigenvalue of the Fréchet derivative $F_u(r_0,0)$. To this end, we first consider the eigenvalue problem for the linearization of (2.4)–(2.6):

$$\begin{aligned}
&\sigma(\Delta\mathbf{u}-\nabla p)+(r\sigma\theta-s\sigma\zeta)\mathbf{e}=\lambda\mathbf{u},\\
&\Delta\theta+w=\lambda\theta,\\
&\tau\Delta\zeta+w=\lambda\zeta,\\
&\nabla\cdot\mathbf{u}=0,\\
&\left.\frac{\partial u}{\partial z}\right|_{z=0,1} = \left.\frac{\partial v}{\partial z}\right|_{z=0,1} = w|_{z=0,1} = \theta|_{z=0,1} = \zeta|_{z=0,1} = 0.
\end{aligned} \tag{3.2}$$

As in the Bénard problem, we seek solutions corresponding to a regular pattern of convection cells. Since (2.4)–(2.6) and the domain Ω_1 are invariant under translations in the xy-plane, we can require that the fields $u=(u,\theta,\zeta)$ are doubly periodic in x and y:

$$u\left(x+\frac{2\pi}{\alpha},y,z\right) = u\left(x,y+\frac{2\pi}{\beta},z\right) = u(x,y,z) \quad \text{for all } \mathbf{x}\in\Omega_1, \tag{3.3}$$

for some nonnegative numbers α, β satisfying $\alpha^2+\beta^2\neq 0$. If we expand the fields in Fourier series

$$u(x,y,z) = \sum_{j,k=-\infty}^{\infty} u_{jk}(z)e^{i(j\alpha x+k\beta y)},\quad u_{-j,-k}(z)=\overline{u_{jk}(z)}, \tag{3.4}$$

and substitute (3.4) into (3.2), we obtain an infinite system of ordinary differential equations

$$
\begin{aligned}
&L_{jk}u_{jk} - ij\alpha p_{jk} = \sigma^{-1}\lambda u_{jk},\\
&L_{jk}v_{jk} - ik\beta p_{jk} = \sigma^{-1}\lambda v_{jk},\\
&L_{jk}w_{jk} - p^{*}_{jk} + r\theta_{jk} - s\zeta_{jk} = \sigma^{-1}\lambda w_{jk},\\
&L_{jk}\theta_{jk} + w_{jk} = \lambda\theta_{jk},\\
&\tau L_{jk}\zeta_{jk} + w_{jk} = \lambda\zeta_{jk},\\
&ij\alpha u_{jk} + ik\beta v_{jk} + w'_{jk} = 0,\\
&u'_{jk}\big|_{z=0,1} = v'_{jk}\big|_{z=0,1} = w_{jk}\big|_{z=0,1} = \theta_{jk}\big|_{z=0,1} = \zeta_{jk}\big|_{z=0,1} = 0
\end{aligned}
\tag{3.5}
$$

for each $(j,k)\in\mathbb{Z}\times\mathbb{Z}$, where $' = d/dz$, $L_{jk} = d^2/dz^2 - \omega_{jk}^2$ and $\omega_{jk}^2 = j^2\alpha^2 + k^2\beta^2$. These equations can be reduced to a single equation for $w_{jk}(z)$,

$$
\begin{aligned}
&\left[(\tau L_{jk}-\lambda)(L_{jk}-\lambda)(L_{jk}-\sigma^{-1}\lambda)L_{jk} + \omega_{jk}^2 r(\tau L_{jk}-\lambda) - \omega_{jk}^2 s(L_{jk}-\lambda)\right] w_{jk} = 0,\\
&w_{jk}\big|_{z=0,1} = w''_{jk}\big|_{z=0,1} = w^{(4)}_{jk}\big|_{z=0,1} = w^{(6)}_{jk}\big|_{z=0,1} = 0
\end{aligned}
\tag{3.6}
$$

for each $(j,k)\in\mathbb{Z}\times\mathbb{Z}$. Because of the boundary conditions, we can expand $w_{jk}(z)$ in a Fourier sine series

$$
w_{jk}(z) = \sum_{l=1}^{\infty} w_{jkl}\sin l\pi z \tag{3.7}
$$

for each $(j,k)\in\mathbb{Z}\times\mathbb{Z}$. Substituting (3.7) into (3.6), we obtain an algebraic equation for λ

(3.8)

$$
\lambda^3 + (\sigma+\tau+1)\gamma_{jkl}^2\lambda^2 + \left[(\sigma+\tau+\sigma\tau)\gamma_{jkl}^4 - \sigma\omega_{jk}^2\gamma_{jkl}^{-2}(r-s)\right]\lambda + \sigma\tau\gamma_{jkl}^6 + \sigma\omega_{jk}^2(s-\tau r) = 0
$$

for each $(j,k)\in\mathbb{Z}\times\mathbb{Z}$ and $l = 1,2,\cdots$, where $\gamma_{jkl}^2 = \omega_{jk}^2 + l^2\pi^2$. Thus the eigenvalues λ of the problem (3.2)–(3.3) are the roots of the cubic polynomials (3.8).

The location of the roots of (3.8) was studied by Baines and Gill [2], and a convenient summary is presented in [7, §2]. For certain values of r, s, σ and τ both zero and purely imaginary roots of (3.8) occur. For now we only need the fact that $\lambda = 0$ is a root of (3.8) if and only if

$$
r = \frac{s}{\tau} + \frac{\left(\omega_{jk}^2 + l^2\pi^2\right)^3}{\omega_{jk}^2} \tag{3.9}
$$

for each $(j,k)\in\mathbb{Z}\times\mathbb{Z}$. Furthermore, $\lambda = 0$ is a simple root of (3.8) provided $\tau \geqq 1$, or if $0 < \tau < 1$ and

$$
s \neq \frac{\tau^2(\sigma+1)\gamma_{jkl}^6}{\sigma(1-\tau)\omega_{jk}^2}. \tag{3.10}
$$

The case $\tau > 1$ corresponds to interchanging the roles of heat and solute to the case in which the temperature gradient is stabilizing and the solute gradient is destabilizing. This is the "fingering" regime. In the rest of this paper, we will restrict τ to $0 < \tau < 1$, the "diffusive" regime.

To determine the multiplicity of $\lambda = 0$ as an eigenvalue of (3.2)–(3.3), we note that for $\lambda = 0$ the boundary-value problem (3.6) reduces to

$$\left[L_{jk}^3 + \omega_{jk}^2\left(r - \frac{s}{\tau}\right)\right] w_{jk} = 0, \tag{3.11}$$
$$w_{jk}|_{z=0,1} = w_{jk}^{(2)}|_{z=0,1} = w_{jk}^{(4)}|_{z=0,1} = 0$$

for each $(j,k) \in \mathbb{Z} \times \mathbb{Z}$. Then (3.11) is equivalent to the integral equation

$$w_{jk}(z) = \mu_{jk} \int_0^1 K_{jk}(z,\hat{z}) w_{jk}(\hat{z})\, d\hat{z} \tag{3.11'}$$

for each (j,k) where $\mu_{jk} = \omega_{jk}^2(r - s/\tau)$, $K_{jk}(z,\hat{z})$ is the composition $K_{jk} = G_{jk} \circ G_{jk} \circ G_{jk}$, and $G_{jk}(z,\hat{z})$ is the Green's function for the regular Sturm–Liouville operator $(-L_{jk})$ defined by

$$(-L_{jk}) w_{jk} = \left(-\frac{d^2}{dz^2} + \omega_{jk}^2\right) w_{jk}, \qquad w_{jk}|_{z=0,1} = 0.$$

Since $(-L_{jk})$ has a positive spectrum, G_{jk} is an oscillating kernel [11, p. 538]. The composition K_{jk} of oscillating kernels is also an oscillating kernel, and one of the consequences of the theory of oscillating kernels is that the eigenvalues of the integral equation (3.11′) are simple and satisfy

$$0 < \mu_{jk1} < \mu_{jk2} < \cdots \to \infty$$

for each $(j,k) \in \mathbb{Z} \times \mathbb{Z}$ [10, Vol. I, pp. 251–255]. In fact, we have

$$\mu_{jkl} = \left(\omega_{jk}^2 + l^2\pi^2\right)^3, \qquad l = 1,2,\cdots \tag{3.12}$$

for each $(j,k) \in \mathbb{Z} \times \mathbb{Z}$. Hence, the smallest eigenvalue of (3.11) corresponds to $l = 1$. Now, $r = r(\omega_{jk}^2)$ given by (3.9) is a convex function of ω_{jk}^2 with a unique minimum when $l = 1$ at $\omega_{jk}^2 = \pi^2/2$, for which

$$r\left(\frac{\pi^2}{2}\right) = r_0 \equiv \frac{s}{\tau} + \frac{27\pi^4}{4}. \tag{3.13}$$

Thus $r = r_0$ is the smallest possible value for r which $\lambda = 0$ is an eigenvalue of (3.2)–(3.3). Moreover, this value of r corresponds to a one-dimensional subspace of nontrivial solutions of the integral equation (3.11) and hence of the system (3.5) for each (j,k) such that $\omega_{jk}^2 = \pi^2/2$ (see [9] or [17] for the application of oscillating kernel theory to the Bénard problem).

If we choose α and β such that

$$\alpha^2 + \beta^2 = \frac{\pi^2}{2}, \tag{3.14}$$

then the multiplicity of $\lambda = 0$ as an eigenvalue of (3.2)–(3.3) is equal to the number of grid points $(j,k) \in \mathbb{Z} \times \mathbb{Z}$ on the ellipse $x^2\alpha^2 + y^2\beta^2 = \pi^2/2$ in the xy-plane. This number is at least four if both α and β are nonzero, or two if either α or β vanishes. Thus, to ensure that $\lambda = 0$ is a simple eigenvalue of (3.2)–(3.3), we must further restrict the space of allowable solutions.

If we require that the solutions of (3.2)–(3.3) are covariant with respect to rotations in the xy-plane,

$$\mathbf{u}(g\mathbf{x}) = g\mathbf{u}(\mathbf{x}), \qquad (\theta,\zeta,p)(g\mathbf{x}) = (\theta,\zeta,p)(\mathbf{x}) \tag{3.15}$$

for all $x \in \Omega_1$ where g is a rotation of the form

$$g = \begin{bmatrix} \cos & -\sin\phi & 0 \\ \sin\phi & \cos\phi & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad 0 \leqq \phi < 2\pi,$$

then the only possible ϕ for which there are nontrivial solutions of (3.2) also satisfying (3.3) are integer multiples of $\phi = 2\pi/k$, $k = 1,2,3,4$ or 6 [14]. We consider four cases which yield function spaces, with respect to which $\lambda = 0$ is a simple eigenvalue of (3.2)–(3.3) (consult [13] for details).

a) *Rolls.* We take $\alpha^2 = \pi^2/2$, $\beta^2 = 0$ in (3.3), and $\phi = \pi$ in (3.15). The flow is two-dimensional with $\mathbf{u}(\mathbf{x}) = (u(x,z), w(x,z))$. Rolls admit a Fourier series expansion of the form

$$u(x,z) = \sum_{j=0}^{\infty} \begin{bmatrix} u_j(z)\sin j\alpha x \\ w_j(z)\cos j\alpha x \\ \theta_j(z)\cos j\alpha x \\ \zeta_j(z)\cos j\alpha x \end{bmatrix} \tag{3.16}$$

and the eigenspace of roll-like solutions of (3.2)–(3.3) for $\lambda = 0$ when $r = r_0$ is the span of

$$u_0(x,z) = \begin{bmatrix} -\dfrac{2\alpha}{\pi}\sin\alpha x\cos\pi z \\ \cos\alpha x\sin\pi z \\ \dfrac{2}{3\pi^2}\cos\alpha x\sin\pi z \\ \dfrac{2}{3\pi^2\tau}\cos\alpha x\sin\pi z \end{bmatrix}. \tag{3.17}$$

b) *Rectangles.* We take $\alpha^2 + \beta^2 = \pi^2/2$, $\alpha, \beta > 0$, $\beta \neq \sqrt{m^2 - 1}\,\alpha$ for all $m \in \mathbb{N}$ in (3.3), and $\phi = \pi$ in (3.15). In addition, we require the invariance of $w(x,y,z)$ under the reflection $y \to -y$. Rectangles admit a Fourier series expansion of the form (cf. [17])

$$u(x,y,z) = \sum_{j,k=0}^{\infty} \begin{bmatrix} u_{jk}(z)\sin j\alpha x\cos k\beta y \\ v_{jk}(z)\cos j\alpha x\sin k\beta y \\ w_{jk}(z)\cos j\alpha x\cos k\beta y \\ \theta_{jk}(z)\cos j\alpha x\cos k\beta y \\ \zeta_{jk}(z)\cos j\alpha x\cos k\beta y \end{bmatrix}, \tag{3.18}$$

and the eigenspace of rectangular solutions of (3.2)–(3.3) for $\lambda=0$ when $r=r_0$ is the span of

$$u_0(x,y,z)=\begin{bmatrix} -\frac{2\alpha}{\pi}\sin\alpha x\cos\beta y\cos\pi z \\ -\frac{2\beta}{\pi}\cos\alpha x\sin\beta y\cos\pi z \\ \cos\alpha x\cos\beta y\sin\pi z \\ \frac{2}{3\pi^2}\cos\alpha x\cos\beta y\sin\pi z \\ \frac{2}{3\pi^2\tau}\cos\alpha x\cos\beta y\sin\pi z \end{bmatrix}, \tag{3.19}$$

c) *Squares.* We take $\alpha^2=\pi^2/4$, $\beta=\alpha$ in (3.3), and $\phi=\pi/2$ in (3.15). Then the Fourier series expansion and eigenfunction are as for rectangles, with $\beta=\alpha$.

d) *Hexagons.* Take $\alpha^2=\pi^2/8$, $\beta=\sqrt{3}\,\alpha$ in (3.3) and $\phi=2\pi/3$ in (3.15). In addition, $w(x,y,z)$ is required to be invariant under $y\to -y$. Hexagons admit a Fourier series expansion of the form (cf. [9])

$$\begin{aligned} \mathbf{u}(x,y,z)&=\frac{1}{3}\sum\left\{\mathbf{u}(j,z)e^{i\alpha(\hat{j}\cdot\hat{x})}+g^{-1}\mathbf{u}(\hat{j},z)e^{i\alpha(\hat{j}\cdot\hat{g}\hat{x})}+g\mathbf{u}(\hat{j},z)e^{i\alpha(\hat{j}\cdot\hat{g}^{-1}\hat{x})}\right\}, \\ \begin{bmatrix}\theta(x,y,z)\\ \zeta(x,y,z)\end{bmatrix}&=\frac{1}{3}\sum\begin{bmatrix}\theta(\hat{j},z)\\ \zeta(\hat{j},z)\end{bmatrix}\left\{e^{i\alpha(\hat{j}\cdot\hat{x})}+e^{i\alpha(\hat{j}\cdot\hat{g}\hat{x})}+e^{i\alpha(\hat{j}\cdot\hat{g}^{-1}\hat{x})}\right\}, \end{aligned} \tag{3.20}$$

where $\hat{j}=(j,\sqrt{3}\,k)$, the sums are over all j, k either both even or both odd, $\hat{x}=(x,y)$,

$$g=\left[\begin{array}{ccc} & \hat{g} & \begin{matrix}0\\ \\ 0\end{matrix} \\ \hline 0 & 0 & 1 \end{array}\right],$$

$\mathbf{u}(\hat{g}\hat{j},z)=g\mathbf{u}(\hat{j},z)$ and $(\theta,\zeta)(\hat{g}\hat{j},z)=(\theta,\zeta)(\hat{j},z)$ for all admissible $\hat{j}$, and the eigenspace of hexagonal eigenfunctions for $\lambda=0$ when $r=r_0$ is the span of

$$u_0(x,y,z)=\left\{\begin{bmatrix} \frac{i\pi}{4\alpha} \\ \frac{i\sqrt{3}\,\pi}{4\alpha} \\ 1 \\ \frac{2}{3\pi^2} \\ \frac{2}{3\pi^2\tau}\end{bmatrix}e^{i\alpha(x+\sqrt{3}y)}+\begin{bmatrix} \frac{i\pi}{4\alpha} \\ -\frac{i\sqrt{3}\,\pi}{4\alpha} \\ 1 \\ \frac{2}{3\pi^2} \\ \frac{2}{3\pi^2\tau}\end{bmatrix}e^{i\alpha(x-\sqrt{3}y)}+\begin{bmatrix} -\frac{i\pi}{2\alpha} \\ 0 \\ 1 \\ \frac{2}{3\pi^2} \\ \frac{2}{3\pi^2\tau}\end{bmatrix}e^{-i2\alpha x}+\text{c.c.}\right\}, \tag{3.21}$$

where c.c. denotes the complex conjugates of the preceding terms.

In each of the above cases a)–d), we say that a function *admits periodicity* Ω if it satisfies the symmetries corresponding to each of the cellular flows a)–d). Such functions are spatially periodic, with fundamental domain of spatial periodicity Ω. For rolls,

$$\Omega = \left\{ (x,z) : 0 < x < \frac{2\pi}{\alpha},\ 0 < z < 1 \right\};$$

for rectangles and squares,

$$\Omega = \left\{ (x,y,z) : 0 < x < \frac{2\pi}{\alpha},\ 0 < y < \frac{2\pi}{\beta},\ 0 < z < 1 \right\};$$

for hexagons, $\Omega = C \times (0,1)$, where C is the region in $\mathbb{R}^2$ enclosed by the six lines $y = \pm\pi/\sqrt{3}\,\alpha$, $y + \sqrt{3}\,x = \pm 2\pi/\sqrt{3}\,\alpha$, $y - \sqrt{3}\,x = \pm 2\pi/\sqrt{3}\,\alpha$ in Fig. 2.

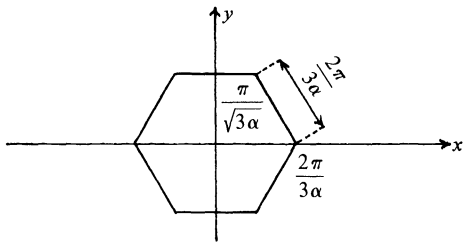


FIG. 2. *The region C.*

The adjoint problem. Integration by parts yields the adjoint eigenvalue problem to (3.2),

$$\begin{aligned}
&\sigma(\Delta \mathbf{u}^* - \nabla p^*) + (\theta^* + \zeta^*)\mathbf{e} = \bar{\lambda}\mathbf{u}^*, \\
&\Delta\theta^* + r\sigma w^* = \bar{\lambda}\theta^*, \\
&\tau\Delta\zeta^* - s\sigma w^* = \bar{\lambda}\zeta^*, \\
&\nabla \cdot \mathbf{u}^* = 0, \\
&\left.\frac{\partial u^*}{\partial z}\right|_{z=0,1} = \left.\frac{\partial v^*}{\partial z}\right|_{z=0,1} = w^*|_{z=0,1} = \theta^*|_{z=0,1} = \zeta^*|_{z=0,1} = 0.
\end{aligned} \tag{3.22}$$

We observe that $(\mathbf{u}_0, p_0, \theta_0, \zeta_0)$ solves (3.2) when $r = r_0$ and $\lambda = 0$ if and only if $(\mathbf{u}_0^*, p_0^*, \theta_0^*, \zeta_0^*) = N(\mathbf{u}_0, p_0, r_0\sigma\theta_0, -s\sigma\zeta_0)$, $N \neq 0$ solves (3.22) when $r = r_0$ and $\bar{\lambda} = 0$. Thus for each of the cellular flows a)–d) the adjoint problem (3.22) also has a one-dimensional null space.

Abstract formulation. Let $n = 2$ (for rolls), or 3 (for rectangles, squares or hexagons) and let $\Omega \in \mathbb{R}^n$ be the fundamental domain of spatial periodicity for functions admitting periodicity Ω. Also, let $\Omega_1 = \mathbb{R}^{n-1} \times (0,1)$, and define the following vector spaces of smooth functions from $\bar{\Omega}$ or Ω into $\mathbb{R}^m$

$$C^{\infty,\#}(\bar{\Omega}, \mathbb{R}^m) = \left\{ u \in C^\infty(\bar{\Omega}_1, \mathbb{R}^m) : u \text{ admits periodicity } \Omega \right\},$$

$$C_0^{\infty,\#}(\Omega, \mathbb{R}^m) = \left\{ u \in C^\infty(\Omega_1, \mathbb{R}^m) : u \text{ admits periodicity } \Omega,\ \operatorname{supp} u \subset \Omega_1 \right\},$$

for $m = 1, 2, \cdots$, and

$$\mathscr{J} = \left\{ \mathbf{u} \in C_0^{\infty,\#}(\Omega, \mathbb{R}^n) : \nabla \cdot \mathbf{u} = 0 \right\},$$

$$\mathscr{V} = \left\{ \mathbf{u} \in C^{\infty,\#}(\bar{\Omega}, \mathbb{R}^n) : \nabla \cdot \mathbf{u} = 0,\ (\mathbf{u} \cdot \mathbf{N})|_{z=0,1} = 0 \right\},$$

$$\mathscr{W} = \mathscr{V} \times C_0^{\infty,\#}(\Omega, \mathbb{R}^2),$$

where $\mathbf{N}$ is the unit outward normal vector on the boundary $z=0,1$. We define the Hilbert spaces

$H^0(\Omega,\mathbb{R}^m)=$ the completion of $C_0^{\infty,\#}(\Omega,\mathbb{R}^m)$ in the norm $\|\cdot\|_0$ associated with the inner product $(u,v)_0=\sum_{i=1}^m \int_\Omega u_i v_i dx$,

$H^k(\Omega,\mathbb{R}^m)=$ the completion of $C^{\infty,\#}(\bar{\Omega},\mathbb{R}^m)$ in the norm $\|\cdot\|_k$ associated with the inner product $(u,v)_k=\sum_{i=1}^m \sum_{|\nu|\le k} \int_\Omega D^\nu u_i D^\nu v_i dx$,

where $k=1,2,\cdots,\nu=(\nu_1,\cdots,\nu_n)$ is an n-tuple of nonnegative integers, D^ν is the partial derivative $\partial^{|\nu|}/\partial x_1^{\nu_1}\cdots\partial x_n^{\nu_n}$ of order $|\nu|=\nu_1+\cdots+\nu_n$,

$H_0^k(\Omega,\mathbb{R}^m)=$ the closure of $C_0^{\infty,\#}(\Omega,\mathbb{R}^m)$ in $H^k(\Omega,\mathbb{R}^m)$,

$J=$ the closure of $\mathscr{J}$ in $H^0(\Omega,\mathbb{R}^n)$,

$V=$ the closure of $\mathscr{V}$ in $H^1(\Omega,\mathbb{R}^n)$,

$W=$ the closure of $\mathscr{W}$ in $H^1(\Omega,\mathbb{R}^{n+2})$.

We note that $W=V\times H_0^1(\Omega,\mathbb{R})\times H_0^1(\Omega,\mathbb{R})$, and that $\mathscr{V}$ is dense in J. Finally we define the following Hilbert spaces

$$X=\left\{u=(\mathbf{u},\theta,\zeta)\in W\cap H^2(\Omega,\mathbb{R}^{n+2}):\ \left.\frac{\partial u}{\partial z}\right|_{z=0,1}=\left.\frac{\partial v}{\partial z}\right|_{z=0,1}=0\right\},$$

$$Y=J\times H^0(\Omega,\mathbb{R})\times H^0(\Omega,\mathbb{R})$$

where $\partial u/\partial z|_{z=0,1}=\gamma_{\mathbf{N}}(\nabla u)$ and $\gamma_{\mathbf{N}}$ is the trace operator extending $\gamma_{\mathbf{N}}\mathbf{u}=(\mathbf{u}\cdot\mathbf{N})|_{z=0,1}$ defined for smooth vector fields $\mathbf{u}$ [18].

We have the following properties [8]:

a) $H^0(\Omega,\mathbb{R}^n)=J\oplus J^\perp$, the topological direct sum of closed subspaces, where $J^\perp$ is the orthogonal complement of J in $H^0(\Omega,\mathbb{R}^n)$,

$$J^\perp=\{\nabla p: p\in H^1(\Omega,\mathbb{R})\}.$$

b) If Π denotes the orthogonal projection of $H^0(\Omega,\mathbb{R}^n)$ onto J, then Π is a bounded linear mapping from $H^k(\Omega,\mathbb{R}^n)$ into $H^k(\Omega,\mathbb{R}^n)\cap J$ for $k=0,1,2,\cdots$. We use Π to eliminate the pressure term in the equations, and formulate (2.6) as

$$\begin{aligned}\frac{d}{dt}\mathbf{u}&=\sigma\Pi[\Delta\mathbf{u}+(r\theta-s\zeta)\mathbf{e}]-\Pi(\mathbf{u}\cdot\nabla)\mathbf{u},\\ \frac{d}{dt}\theta&=\Delta\theta+w-\mathbf{u}\cdot\nabla\theta,\\ \frac{d}{dt}\zeta&=\tau\Delta\zeta+w-\mathbf{u}\cdot\nabla\zeta,\end{aligned}\tag{3.23}$$

for $(\mathbf{u},\theta,\zeta)\in X$. If we define

$$L(r)u=(\sigma\Pi[\Delta\mathbf{u}+(r\theta-s\zeta)\mathbf{e}],\ \Delta\theta+w,\ \tau\Delta\zeta+w)\tag{3.24}$$

for $u=(\mathbf{u},\theta,\zeta)\in X$ and $r>0$, then $L(r)$ is a bounded linear mapping from X into Y for all $r>0$. Define

$$M(u_1,u_2)=(-\Pi(\mathbf{u}_1\cdot\nabla)\mathbf{u}_2,\ -\mathbf{u}_1\cdot\nabla\theta_2,\ -\mathbf{u}_1\cdot\nabla\zeta_2)\tag{3.25}$$

for $u_i=(\mathbf{u}_i,\theta_i,\zeta_i)\in X$, $i=1,2$. Then M is a bounded bilinear mapping from $X\times X$ into Y [8].

We now let $F:(0,\infty)\times X\to Y$ be defined by

$$F(r,u)=L(r)u+M(u,u). \tag{3.26}$$

Then F is analytic, $F(r,0)=0$ for all $r\in(0,\infty)$, $F_u(r,0)=L(r)$ for all $r\in(0,\infty)$ and the null space $N(L(r_0))$ is a one-dimensional subspace of the space X of functions admitting periodicity Ω for each choice of Ω. To show that the range $R(L(r_0))$ has codimension one, we use the following lemma, which is proved in the appendix.

LEMMA 1. $R(L(r_0))=N(L(r_0)^*)^{\perp}$.

Since the adjoint problem has a one-dimensional null space $N(L(r_0)^*)$, we now have satisfied all the hypotheses of Theorem 1 except (iii). To verify this last hypothesis, it suffices to check that

$$\left(\Pi(\theta_0\mathbf{e}),u_0^*\right)_J\neq 0, \tag{3.27}$$

where $(\mathbf{u}_0,\theta_0,\zeta_0)$ spans $N(L(r_0))$ and $(\mathbf{u}_0^*,\theta_0^*,\zeta_0^*)=(\sigma^{-1}\mathbf{u}_0,\mathrm{r}_0\theta_0,-\mathrm{s}\zeta_0)$ spans $N(L(r_0)^*)$. Since $\mathbf{u}_0^*\in J$, and hence $\nabla\cdot\mathbf{u}_0^*=0$, (3.27) is equivalent to

$$\int_\Omega\theta_0 w_0\,dx\neq 0. \tag{3.28}$$

Since $w_0=-\Delta\theta_0$, θ_0 satisfies periodicity Ω, and $\theta_0|_{z=0,1}=0$, we have $\int_\Omega\theta_0 w_0\,dx=\int_\Omega|\nabla\theta_0|^2\,dx>0$, and hence (3.28) is satisfied for each cellular structure: rolls, rectangles, squares or hexagons.

Finally, we take $Z=N(L(r_0))^{\perp}$. Then Theorem 1 establishes the existence of a bifurcating curve $(r(\varepsilon),u(\varepsilon))$ of nontrivial equilibrium solutions near $(r_0,0)$. For each cellular structure, the bifurcation point corresponds to the same critical value r_0 given by (3.13).

4. Linearized stability of the bifurcating solutions. In this section we investigate the linearized stability of the bifurcating branch of nontrivial equilibrium solutions $(r(\varepsilon),u(\varepsilon))$, whose existence was proven in §3. This bifurcating solution is stable (resp. unstable) in the linearized sense if the small real eigenvalue $\tilde\lambda(\varepsilon)$ of the linearization $F_u(r(\varepsilon),u(\varepsilon))$ is negative (resp. positive) near $\varepsilon=0$. For Navier–Stokes type systems such as the one we use, Iooss [8] has shown that linearized stability implies the asymptotic stability of $u(\varepsilon)$ with respect to small initial perturbations.

We first summarize the linearized stability theory of Crandall and Rabinowitz [5] and then indicate how to use power series expansions to determine the sign of $\tilde\lambda(\varepsilon)$ near $\varepsilon=0$. We then present the results of our calculations.

Linearized stability. Before stating the main theorem on linearized stability, we first define the concept of a K-simple eigenvalue.

DEFINITION. Let X and Y be real Banach spaces, and let $L,K\in\mathcal{B}(X,Y)$. Then $\lambda\in\mathbb{R}$ is a *K-simple eigenvalue* of L if

i) $\dim N(L-\lambda K)=\operatorname{codim}R(L-\lambda K)=1$, and

ii) $Ku_0\notin R(L-\lambda K)$, where $N(L-\lambda K)=\operatorname{span}\{u_0\}$.

The following result shows that the simple zero eigenvalues of $F_u(r_0,0)$ in Theorem 1 are associated with a small eigenvalue $\tilde\lambda(\varepsilon)$ of the linearization $F_u(r(\varepsilon),u(\varepsilon))$ about the bifurcating branch. Let F, Z, u_0 be as in Theorem 1 and let $r(\varepsilon)$, $u(\varepsilon)=\varepsilon u_0+\varepsilon z(\varepsilon)$ be as supplied by the theorem.

LEMMA 2. [5, p. 165]. *Let $K\in\mathcal{B}(X,Y)$ and suppose zero is a K-simple eigenvalue of $F_u(r,0)$. Then there exist open intervals $\hat{\mathscr{I}},\hat{\mathscr{J}}$ with $r_0\in\hat{\mathscr{I}}$, $D\in\hat{\mathscr{J}}$ and smooth functions*

$$\lambda:\hat{\mathscr{I}}\to\mathbb{R},\quad\tilde\lambda:\hat{\mathscr{J}}\to\mathbb{R},\quad v:\hat{\mathscr{I}}\to X,\quad\tilde v:\hat{\mathscr{J}}\to X$$

such that

$$F_u(r,0)v(r)=\lambda(r)Kv(r) \quad \text{for } r\in\hat{\mathscr{I}},$$
$$F_u(r(\varepsilon),u(\varepsilon))\tilde{v}(\varepsilon)=\tilde{\lambda}(\varepsilon)K\tilde{v}(\varepsilon) \quad \text{for } \varepsilon\in\hat{\mathscr{I}}.$$

Moreover, $\lambda(r_0)=\tilde{\lambda}(0)=0$, $v(r_0)=\tilde{v}(0)=u_0$, $v(r)-u_0\in Z$, $\tilde{v}(\varepsilon)-u_0\in Z$.

The main result of [5] is Theorem 1.16.

THEOREM 2 [5, p. 165]. *Let the assumptions of Theorem* 1 *and Lemma* 2 *hold, and let* λ, $\tilde{\lambda}$ *be the functions supplied by Lemma* 2. *Then* $\lambda'(r_0)\neq 0$ *and*

$$\lim_{\substack{\varepsilon\to 0\\ \tilde{\lambda}(\varepsilon)\neq 0}}\left[-\frac{\varepsilon r'(\varepsilon)\lambda'(r_0)}{\tilde{\lambda}(\varepsilon)}\right]=1.$$

Moreover, there is a constant C *such that* $\|u'(\varepsilon)-\tilde{v}(\varepsilon)\|\leqq C\min\{|\varepsilon r'(\varepsilon)|, |\tilde{\lambda}(\varepsilon)|\}$ *near* $\varepsilon=0$.

In what follows we take K to be the continuous injection of X into Y, $Ku=u$. By Theorem 2, the sign of $\tilde{\lambda}(\varepsilon)$ near $\varepsilon=0$ is the same as the sign of $-\varepsilon r'(\varepsilon)\lambda'(r_0)$ near $\varepsilon=0$, provided $\tilde{\lambda}(\varepsilon)\neq 0$. To determine the sign of $-\varepsilon r'(\varepsilon)\lambda'(r_0)$, we use power series expansions.

Power series expansions. Since $F(r,u)$ is analytic in r and u, we can expand $r(\varepsilon)$ and $u(\varepsilon)$ in power series

$$\begin{aligned} r(\varepsilon)&=r_0+\varepsilon r_1+\varepsilon^2 r_2+\cdots,\\ u(\varepsilon)&=\varepsilon(u_0+\varepsilon u_1+\cdots). \end{aligned} \tag{4.1}$$

Substituting the power series (4.1) into $F(r(\varepsilon), u(\varepsilon))=0$, taking the inner product with u_0^* and collecting terms we obtain at order ε^2

$$r_1=-\frac{\big(M(u_0,u_0),u_0^*\big)}{\big(F_{ru}(r_0,0)u_0,u_0^*\big)}. \tag{4.2}$$

If $r_1=0$, then at order ε^3 we get

$$r_2=-\frac{\big(M(u_0,u_1)+M(u_1,u_0),u_0^*\big)}{\big(F_{ru}(r_0,0)u_0,u_0^*\big)}, \tag{4.3}$$

where $F_u(r_0,0)u_1=-M(u_0,u_0)$. If $r_2\neq 0$, then $\tilde{\lambda}(\varepsilon)\neq 0$ near $\varepsilon=0$ (expand $\tilde{\lambda}(\varepsilon)=\varepsilon\tilde{\lambda}_1+\varepsilon^2\tilde{\lambda}_2+\cdots$, $\tilde{v}(\varepsilon)=u_0+\varepsilon\tilde{v}_1+\cdots$ and substitute into $F_u(r(\varepsilon),u(\varepsilon))\tilde{v}(\varepsilon)=\tilde{\lambda}(\varepsilon)u(\varepsilon)$) and by Theorem 2, $\tilde{\lambda}(\varepsilon)$ has the same sign as $-r_2\lambda'(r_0)$ near $\varepsilon=0$.

Differentiating the equation $F_u(r,0)v(r)=\lambda(r)v(r)$ with respect to r, evaluating at $r=r_0$ and then taking the inner product with u_0^*, we obtain

$$\big(F_{ru}(r_0,0)u_0,u_0^*\big)=\lambda'(r_0)\big(u_0,u_0^*\big). \tag{4.4}$$

In §3 we determined that the left-hand side of (4.4) was strictly positive, and hence the sign of $\lambda'(r_0)$ is the same as the sign of (u_0,u_0^*) provided $(u_0,u_0^*)\neq 0$. For each cellular structure, simple computations show that

$$\begin{aligned} &\big(u_0,u_0^*\big)>0 \quad \text{if } 0<\tau<1 \text{ and } 0<s<s_C,\\ &\big(u_0,u_0^*\big)=0 \quad \text{if } 0<\tau<1 \text{ and } s=s_C,\\ &\big(u_0,u_0^*\big)<0 \quad \text{if } 0<\tau<1 \text{ and } s>s_C, \end{aligned} \tag{4.5}$$

where

$$s_C = \frac{27\pi^4}{4}\left(\frac{\tau^2}{1-\tau}\right)\left(1+\frac{1}{\sigma}\right)$$

(cf. (3.10)). If $0<\tau<1$ and $0<s<s_c$, then the zero eigenvalue of $F_u(r_0,0)$ crosses into the positive complex half-plane as r increases past r_0, and all the other eigenvalues of $F_u(r_0,0)$ are isolated points in the negative complex half-plane. Thus the bifurcating solution $u(\varepsilon)$ will be stable if $\varepsilon r'(\varepsilon)>0$ near $\varepsilon=0$, and unstable if $\varepsilon r'(\varepsilon)<0$ near $\varepsilon=0$. However, if $0<\tau<1$ and $s>s_c$, then the zero eigenvalue of $F_u(r_0,0)$ crosses into the negative complex half-plane and moreover, $F_u(r_0,0)$ has one positive real eigenvalue [7]. Thus $F_u(r(\varepsilon),u(\varepsilon))$ has a positive eigenvalue near $\varepsilon=0$ and the solution $u(\varepsilon)$ is unstable in this case, regardless of the sign of $\varepsilon r'(\varepsilon)$.

Results. Due to the dependence of the eigenfunctions u_0 on z for each cellular structure, $r_1=0$. Thus to determine the stability of the bifurcating solution $u(\varepsilon)$, we must solve $F_u(r_0,0)u_1=-M(u_0,u_0)$ and compute r_2 from (4.3). Then the sign of $\tilde{\lambda}(\varepsilon)$ near $\varepsilon=0$ is the same as the sign of $-r_2\lambda'(r_0)$ and hence for $0<\tau<1$ and $0<s<s_C$, the bifurcating solution $u(\varepsilon)$ is stable if $r_2>0$ and unstable if $r_2<0$.

a) For *rolls* we have

$$u_1=0, \qquad w_1=0,$$

$$\theta_1=-\frac{1}{12\pi^3}\sin 2\pi z, \qquad \zeta_1=-\frac{1}{12\pi^3\tau}\sin 2\pi z,$$

and r_2 has the same sign as $r_0-s/\tau^3=27\pi^4/4-((1-\tau^2)/\tau^3)s$. Thus $r_2>0$ if $0<\tau<1$ and $0<s<s_1\equiv(27\pi^4/4)(\tau^3/(1-\tau^2))$, while $r_2>0$ if $s>s_1$. We note that $s_1<s_C$ so that the branch of convecting rolls bifurcates subcritically as r increases past r_0 when $s_1<s<s_C$ (and when $s>s_C$). This result is well-known (see [10, Vol. II, p. 45] and the references therein) and moreover, the value s_1 is the same as the global nonlinear stability limit of the conduction solution $u=0$ when $r=r_0$, obtained by Joseph [10, Vol. II, pp. 42–46] using energy methods.

b) For *rectangles* we have

$$u_1=-\frac{\pi}{\alpha}w_{202}\sin 2\alpha x\cos 2\pi z,$$

$$v_1=-\frac{\pi}{\beta}w_{022}\sin 2\beta y\cos 2\pi z,$$

$$w_1=(w_{202}\cos 2\alpha x+w_{022}\cos 2\beta y)\sin 2\pi z,$$

$$\theta_1=\left\{-\frac{1}{24\pi^3}+\frac{1}{4(\pi^2+\alpha^2)}\left[w_{202}-\frac{1}{3\pi^3}\left(\frac{\pi^2}{2}-\alpha^2\right)\right]\cos 2\alpha x\right.$$

$$\left.+\frac{1}{4(\pi^2+\beta^2)}\left[w_{022}-\frac{1}{3\pi^3}\left(\frac{\pi^2}{2}-\beta^2\right)\right]\cos 2\beta y\right\}\sin 2\pi z,$$

$$\zeta_1=\left\{-\frac{1}{24\pi^3\tau}+\frac{1}{4\tau(\pi^2+\alpha^2)}\left[w_{202}-\frac{1}{3\pi^3\tau}\left(\frac{\pi^2}{2}-\alpha^2\right)\right]\cos 2\alpha x\right.$$

$$\left.+\frac{1}{4\tau(\tau^2+\beta^2)}\left[w_{022}-\frac{1}{3\pi^3\tau}\left(\frac{\pi^2}{2}-\beta^2\right)\right]\cos 2\beta y\right\}\sin 2\pi z,$$

where

$$w_{202} = -\frac{2\pi}{\sigma} f(\alpha)\left[1-\frac{2}{\pi^2}(\alpha^2-\beta^2)\right]\left[3+\frac{\sigma}{6\pi^2(\pi^2+\alpha^2)}\left(r_0-\frac{s}{\tau^2}\right)\right],$$

$$w_{022} = -\frac{2\pi}{\sigma} f(\beta)\left[1-\frac{2}{\pi^2}(\beta^2-\alpha^2)\right]\left[3+\frac{\sigma}{6\pi^2(\pi^2+\beta^2)}\left(r_0-\frac{s}{\tau^2}\right)\right],$$

$$f(\alpha) = 16\alpha^2(\pi^2+\alpha^2)\left[64(\pi^2+\alpha^2)^3 - 27\pi^4\alpha^2\right]^{-1}.$$

Then r_2 has the same sign as the expression

$$(4.6)\quad \frac{2\sigma}{9\pi^4}\left(r_0-\frac{s}{\tau^3}\right)\left\{4+\frac{\pi^2-2\alpha^2}{\pi^2+\alpha^2}\left[1-\frac{2}{\pi^2}(\alpha^2-\beta^2)\right]\right.$$

$$\left.+\frac{\pi^2-2\beta^2}{\pi^2+\beta^2}\left[1-\frac{2}{\pi^2}(\beta^2-\alpha^2)\right]\right\}$$

$$+\frac{2}{\sigma}\left\{f(\alpha)(\pi^2-2\alpha^2)\left[1-\frac{2}{\pi^2}(\alpha^2-\beta^2)\right]\left[3+\frac{\sigma}{6\pi^2(\pi^2+\alpha^2)}\left(r_0-\frac{s}{\tau^2}\right)\right]^2\right.$$

$$\left.+f(\beta)(\pi^2-2\beta^2)\left[1-\frac{2}{\pi^2}(\beta^2-\alpha^2)\right]\left[3+\frac{\sigma}{6\pi^2(\pi^2+\beta^2)}\left(r_0-\frac{s}{\tau^2}\right)\right]^2\right\}.$$

We observe that $\pi^2-2\alpha^2$, $\pi^2-2\beta^2$, $[1-(2/\pi^2)(\alpha^2-\beta^2)]$, $[1-(2/\pi^2)(\beta^2-\alpha^2)]$, $f(\alpha)$ and $f(\beta)$ are all strictly positive when $0<\alpha^2, \beta^2<\pi^2/2$.

c) For *squares*, $\alpha^2=\beta^2=\pi^2/4$ and the expression (4.6) simplifies to

$$(4.7)\quad \frac{16\sigma}{15\pi^4}\left(r_0-\frac{s}{\tau^3}\right)+\frac{40}{473\sigma}\left[3+\frac{2\sigma}{15\pi^4}\left(r_0-\frac{s}{\tau^2}\right)\right]^2.$$

d) For *hexagons* we have

$$u_1 = \frac{i\pi}{2\alpha}\left\{w_{112}\Phi_{11}^1 + w_{312}\Phi_{31}^1\right\}\cos 2\pi z,$$

$$v_1 = \frac{i\sqrt{3}\,\pi}{2\alpha}\left\{w_{112}\Phi_{11}^2 + w_{312}\Phi_{31}^2\right\}\cos 2\pi z,$$

$$w_1 = \left\{w_{112}\Phi_{11}^3 + w_{312}\Phi_{31}^3\right\}\cos 2\pi z,$$

$$\theta_1 = \left\{-\frac{2}{\pi^3}+\left[\frac{2}{9\pi^2}w_{112}-\frac{4}{9\pi^3}\right]\Phi_{11}^3+\left[\frac{2}{11\pi^2}w_{312}-\frac{4}{33\pi^3}\right]\Phi_{31}^3\right\}\cos 2\pi z,$$

$$\zeta_1 = \left\{-\frac{2}{\pi^3\tau^2}+\left[\frac{2}{9\pi^2\tau}w_{112}-\frac{4}{9\pi^3\tau^2}\right]\Phi_{11}^3+\left[\frac{2}{11\pi^2\tau}w_{312}-\frac{4}{33\pi^3\tau^2}\right]\Phi_{31}^3\right\}\cos 2\pi z,$$

where

$$w_{112} = -\frac{1}{39\pi\sigma}\left[9+\frac{4\sigma}{9\pi^4}\left(r_0-\frac{s}{\tau^2}\right)\right],$$

$$w_{312} = -\frac{33}{625\pi\sigma}\left[3+\frac{4\sigma}{33\pi^4}\left(r_0-\frac{s}{\tau^2}\right)\right],$$

$$\Phi_{11}^1 = e^{i\alpha(x+\sqrt{3}y)}+e^{i\alpha(x-\sqrt{3}y)}-2e^{-2i\alpha x}-e^{i\alpha(-x-\sqrt{3}y)}$$
$$-e^{i\alpha(-x+\sqrt{3}y)}+2e^{2i\alpha x},$$

$$\Phi_{11}^2 = e^{i\alpha(x\sqrt{3}y)}-e^{i\alpha(x-\sqrt{3}y)}-e^{i\alpha(-x-\sqrt{3}y)}+e^{i\alpha(-x+\sqrt{3}y)},$$

$$\Phi_{11}^3 = e^{i\alpha(x+\sqrt{3}y)}+e^{i\alpha(x-\sqrt{3}y)}+e^{-2i\alpha x}+e^{i\alpha(-x-\sqrt{3}y)}$$
$$+e^{i\alpha(-x+\sqrt{3}y)}+e^{2i\alpha x},$$

$$\Phi_{31}^1 = 3e^{i\alpha(3x+\sqrt{3}y)}+3^{i\alpha(3x-\sqrt{3}y)}-3e^{i\alpha(-3x-\sqrt{3}y)}$$
$$-3e^{i\alpha(-x+\sqrt{3}y)}d,$$

$$\Phi_{31}^2 = e^{i\alpha(3x+\sqrt{3}y)}-e^{i\alpha(3x-\sqrt{3}y)}-2e^{-i2\sqrt{3}\alpha y}$$
$$-e^{i\alpha(-3x-\sqrt{3}y)}+e^{i\alpha(-3x+\sqrt{3}y)}+2e^{i2\sqrt{3}\alpha y},$$

$$\Phi_{31}^3 = e^{i\alpha(3x+\sqrt{3}y)}+e^{i\alpha(3x-\sqrt{3}y)}+e^{-i2\sqrt{3}\alpha y}+e^{i\alpha(-3x-\sqrt{3}y)}$$
$$+e^{i\alpha(-3x+\sqrt{3}y)}+e^{i2\sqrt{3}\alpha y}.$$

Then r_2 has the same sign as the expression

$$\frac{120\sigma}{11\pi^4}\left(r_0-\frac{s}{\tau^3}\right)+\frac{62208}{8125\sigma}+\frac{5936}{8125\pi^4}\left(r_0-\frac{s}{\tau^2}\right)+\frac{126848\sigma}{7239375\pi^8}\left(r_0-\frac{s}{\tau^2}\right)^2. \tag{4.8}$$

The linearized stability of the bifurcating three-dimensional cellular solutions—rectangles, squares and hexagons—all show similar dependence on the parameters s, σ and τ. If we substitute

$$r_0-\frac{s}{\tau^2} = -\frac{27\pi^4}{4\sigma}+\frac{1-\tau}{\tau^2}(s_C-s),$$
$$r_0-\frac{s}{\tau^3} = -\frac{27\pi^4(1+\sigma+\tau)}{4\sigma\tau}+\frac{1-\tau^2}{\tau^3}(s_C-s) \tag{4.9}$$

into the expressions (4.6), (4.7) and (4.8) we obtain

(4.10)

$$-\frac{3}{2}\left(\frac{1+\sigma+\tau}{\tau}\right)\left\{4+\frac{\pi^2-2\alpha^2}{\pi^2+\alpha^2}\left[1-\frac{2}{\pi^2}(\alpha^2-\beta^2)\right]+\frac{\pi^2-2\beta^2}{\pi^2+\beta^2}\left[1-\frac{2}{\pi^2}(\beta^2-\alpha^2)\right]\right\}$$
$$+\frac{2}{\sigma}\left\{f(\alpha)(\pi^2-2\alpha^2)\left[1-\frac{2}{\pi^2}(\alpha^2-\beta^2)\right]\left[\frac{15\pi^2+24\alpha^2}{8(\pi^2+\alpha^2)}\right]^2\right.$$
$$\left.+f(\beta)(\pi^2-2\beta^2)\left[1-\frac{2}{\pi^2}(\beta^2-\alpha^2)\right]\left[\frac{15\pi^2+24\beta^2}{8(\pi^2+\beta^2)}\right]^2\right\}+O(s_C-s)$$

for rectangles,

$$-\frac{36}{5}\left(\frac{1+\sigma+\tau}{\tau}\right)+\frac{84}{437\sigma}+O(s_C-s) \tag{4.11}$$

for squares, and

$$-\frac{810}{11}\left(\frac{1+\sigma+\tau}{\tau}\right)+\frac{314892}{89375\sigma}+O(s_C-s) \tag{4.12}$$

for hexagons. In each case the term $O(s_C-s)$ is of the form

$$O(s_C-s)=A(s_C-s)+B(s_C-s)^2, \qquad A,B>0 \tag{4.13}$$

so that, in particular, $\lim_{s\to s_C^-} O(s_C-s)=0$. From (4.6)–(4.8) we see that given any $\sigma>0$ and $0<\tau<1$ the coefficient r_2 is positive if s is near zero, but can be positive or negative as s increases up to s_C, depending on σ and τ. More explicitly, one sees from (4.10)–(4.13) that given $0<\tau<1$ one can choose σ sufficiently close to zero so that $r_2>0$ for all s, $0<s<s_C$; but given $\sigma>0$ one can choose τ sufficiently close to zero so that $r_2<0$ for all s, $s_1+\varepsilon_0<s<s_C$ where $0<\varepsilon_0<s_C-s_1$.

In Figs. 3–6 we have plotted the coefficient r_2 as a function of the size of the solute concentration gradient s for several values of σ and τ, comparing rolls, squares and hexagons. In Fig. 3, for $\sigma=7$ and $\tau=\frac{1}{80}$ (the appropriate values for salt dissolved in water) the cofficient r_2 is positive at $s=0$ and changes sign at a small positive value for s: for rolls r_2 changes sign at $s=0.00128$, while for squares and hexagons r_2 changes sign at slightly higher values $s=0.001311$ (squares), $s=0.001313$ (hexagons). As σ decreases towards 0, the differences become greater. In Fig. 4, for $\sigma=1$ and $\tau=0.1$ and more clearly in Fig. 5, for $\sigma=0.01$ and $\tau=0.1$ one can see that r_2 for squares and hexagons changes sign at higher values of s than does r_2 for rolls. For sufficiently small σ, as in Fig. 6, where $\sigma=0.001$ and $\tau=0.1$, the coefficients r_2 for squares and hexagons remain positive for all s, $0<s<s_C$. In contrast, r_2 for rolls is negative except for s very near $s=0$.

For most systems of physical interest, $0<\tau\ll 1$ and $\sigma\gtrsim 1$ so that the dependence of the coefficient r_2 on s for the three-dimensional cellular solutions is nearly the same as that for roll-like solutions, although r_2 is positive for a slightly larger range of s for the three-dimensional cellular solutions.

The sign of r_2 indicates the directions of bifurcation (supercritical if r_2 is positive, subcritical if r_2 is negative) and the stability of the equilibria on the branch of convective solutions, but only with respect to small perturbations having the same cellular structure. It is possible, for example, that solutions on a supercritical branch of hexagons are stable with respect to small perturbations having the same hexagonal structure, but are unstable with respect to other small perturbations. We discuss this possibility in the next section.

5. Comments on pattern selection. In a recent paper [20], Golubitsky, Swift and Knobloch investigate pattern selection between rolls, hexagons and two other cellular structures in a model of pure thermal (Bénard) convection with the same boundary conditions on the upper and lower surfaces. They treat bifurcation problems that are symmetric with respect to the group preserving doubly periodic functions on a hexagonal lattice. After reducing the problem at the bifurcation point to a six-dimensional phase space, all possible local bifurcation diagrams are found (given certain nondegeneracy conditions) involving rolls, hexagons, regular triangles and a "patchwork quilt"

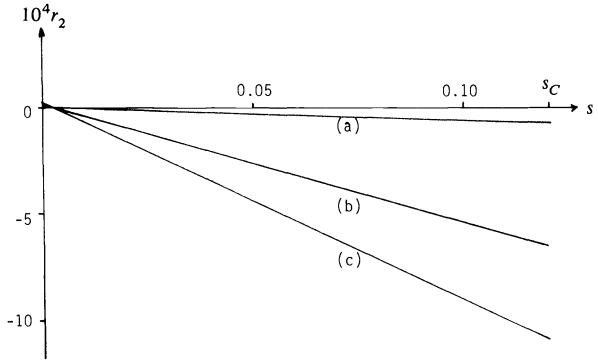


FIG. 3. *The coefficient* r_2 *as a function of* s *when* $\sigma = 7$, $\tau = \frac{1}{80}$ *for* (a) *rolls*, (b) *squares and* (c) *hexagons* ($s_C = 0.119$).

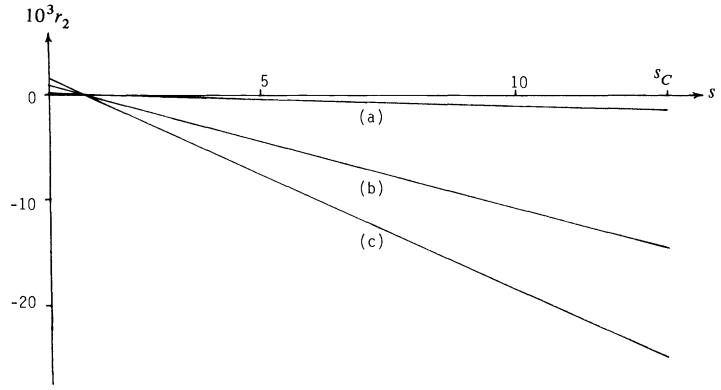


FIG. 4. *The coefficient* r_2 *as a function of* s *when* $\sigma = 1$, $\tau = 0{,}1$ *for* (a) *rolls*, (b) *squares and* (c) *hexagons* ($s_C = 14.6$).

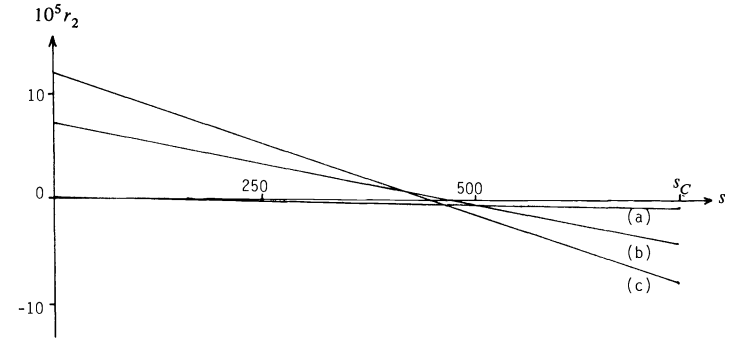


FIG. 5. *The coefficient* r_2 *as a function of* s *when* $\sigma = 0.01$, $\tau = 0.1$ *for* (a) *rolls*, (b) *squares and* (c) *hexagons* ($s_C = 738$).

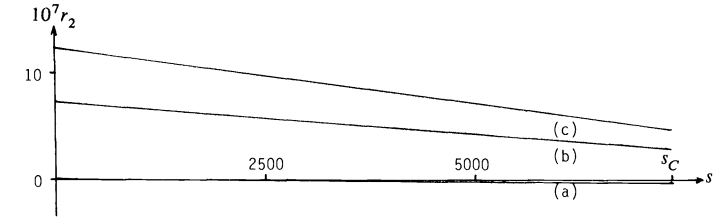


FIG. 6. *The coefficient* r_2 *as a function of* s *when* $\sigma = 0.001$, $\tau = 0.1$ *for* (a) *rolls*, (b) *squares and* (c) *hexagons* ($s_C = 7131$).

of rectangles. Their results are applicable to the convection problem in this paper. In particular, the coefficients r_2 which we have computed for rolls and hexagons can be used to find the coefficients $l_3(0)$ and a of [20]. Provided a certain fifth order term $m_5(0)$ in the reduced bifurcation equations is nonzero, we have the following pattern selection results: (a) If $l_3(0)>0$ and $a<-1$, then rolls are stable in the sense of [20], i.e., near the bifurcation point rolls are stable with respect to small perturbations possessing the symmetries of the hexagonal lattice and midplane reflection; (b) if $l_3(0)<0$ and $a>-\frac{1}{3}$, then either hexagons of regular triangles are stable, depending on the sign of $m_5(0)$; (c) for all other values of $l_3(0)$ and a, solutions near the bifurcation point are unstable even though they may be on a supercritical branch.

The value of the coefficient $m_5(0)$ depends on the results of computations for the coefficient r_4 in (4.1). Although we have not computed r_4, both r_4 and $m_5(0)$ are most likely nontrivial functions of the parameters σ, τ and s, and $m_5(0)\neq 0$ for almost all parameter values $\sigma>0$, $0<\tau<1$, $0<s<s_C$.

In Fig. 7 we have plotted the coefficient a as a function of s, for $\sigma=0.01$ and $\tau=0.1$. We observe the following behavior: for $s=0$ and for s in a small right neighborhood of $s=0$, case (a) holds; and for all other values of s less than s_C, case (c) holds. Assuming that $m_5(0)\neq 0$, this implies that rolls are stable in the sense of [20] for $s=0$ and for s in the small right neighborhood of $s=0$, but for all other values of s less than s_C, neither rolls nor hexagons are stable, even though the branch of hexagons may bifurcate supercritically. For the other values of σ and τ corresponding to Figs. 3, 4 and 6 we observed similar behavior of the coefficients $l_3(0)$ and a as functions of s. In fact, it is easy to show that when $s=0$, and by continuity when s belongs to some right neighborhood of $s=0$, then $l_3(0)>0$ and $a<-1$. The case $s=0$ corresponds to pure thermal (Bénard) convection and this result agrees with that of [21]. We did not find any parameter values yielding case (b) in which stable hexagons are possible.

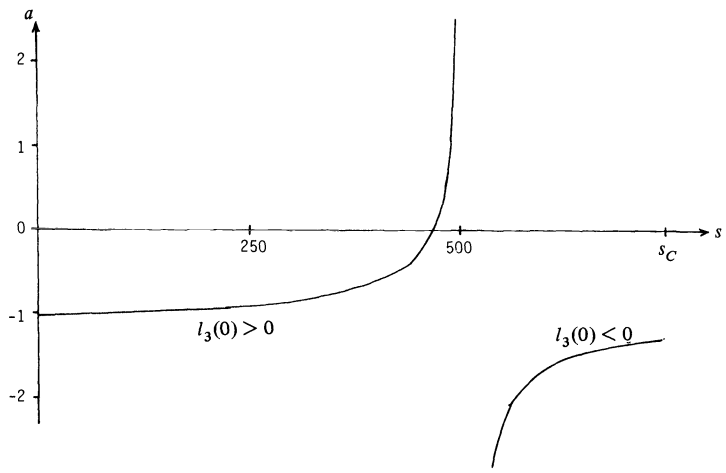


FIG. 7. *The hexagonal lattice coefficient a as a function of s when* $\sigma=0.01$, $\tau=0.1$ ($s_C=738$).

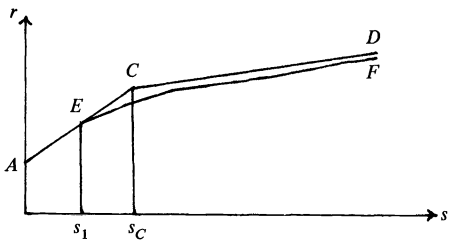
The pattern selection results do not apply to the squares and rectangles for which we have also computed values for r_2. A theory of pattern selection including more general cellular structures (for example, squares) with rolls and hexagons is not yet developed. However, in experiments performed on layers of fluid with finite extent, one usually sees either rolls or hexagons.

6. Conclusion. We have used a bifurcation theorem of Crandall and Rabinowitz to prove the existence of nontrivial equilibrium solutions of the doubly-diffusive convection equations. For fixed parameter values $\sigma > 0$ (Prandtl number), $0 < \tau < 1$ (ratio of solute diffusivity to thermal diffusivity) and $s \neq s_C$ (solute concentration gradient) there exist solutions corresponding to steady cellular convection for r (temperature gradient) in a neighborhood of the critical value r_0.

At the bifurcation point $u = 0$, $r = r_0$, four types of convective solutions (rolls, rectangles, squares and hexgaons) bifurcate supercritically when s is near 0. However, when s is near s_C rolls always bifurcate subcritically but three-dimensional cellular structures bifurcate supercritically if σ is sufficiently close to 0. Solutions on a subcritical branch near the bifurcation point are unstable, but solutions on a supercritical branch may only be stable with respect to a very restricted class of perturbations. Using the results of Golubitsky, Swift and Knobloch [20], we have found that near $s = 0$ rolls are stable with respect to a class of perturbations including both rolls and hexagons (assuming a nondegeneracy condition); but for the parameter values we explored, hexagons are unstable with respect to this class of perturbations even though the hexagon solutions may be on a supercritical branch.

Although we give only local existence and stability results, in a neighborhood of the bifurcation point $r = r_0$, $u = 0$, other methods applied to this problem have provided information on the global behavior of the bifurcating branches of equilibrium solutions. Using energy methods, Joseph studied the global stability of the conduction solution and for $0 < \tau < 1$ gives the following diagram [10, Vol. II, p. 46] (cf. Fig. 1). For (r,s) values above the curve AECD, the conduction solution is unstable (Fig. 8). For (r,s) values below the curve AEF, the conduction solution is globally stable. For (r,s) values between the two curves, the conduction solution is stable with respect to small perturbations, but other equilibrium solutions are possible. In fact, the two-dimensional roll-like solutions bifurcate subcritically ($r_2 < 0$, and hence $r(\varepsilon) < r_0$ near $\varepsilon = 0$) for values of (r,s) on the open line segment EC, and hence these solutions exist for (r,s) in a neighborhood below EC. Since they cannot exist for (r,s) below EF, it is plausible that the branch of solutions in ru-space for fixed s, $s_1 < s < s_C$, "turns back" and regains stability as is suggested by numerical calculations [7] and by the results of a perturbation analysis near $s = s_1$ [16]. The five-dimensional system of ordinary differential equations obtained by model truncation of the two-dimensional flow also has a branch of solutions corresponding to convection which bifurcates subcritically from the conduction solution and possesses a turning point [6]. Thus, it is reasonable to conjecture that this same behavior is repeated in the infinite-dimensional case.

The doubly-diffusive convection equations also admit periodic solutions, which have not been considered here. When $0 < \tau < 1$ and $s > s_C$, the bifurcating equilibrium



AECD: instability limit of conduction solution

AEF: stability limit of conduction solution

$r = r_0$ on AEC.

FIG. 8. *Stability and instability limits of the conduction solution.*

solutions near $r=r_0$, $u=0$ are unstable, but on the curve CD in Figs. 1 and 3, the linearization possesses two conjugate pure imaginary eigenvalues. We therefore expect a Hopf bifurcation of periodic solutions [7]. Furthermore, if $\sigma>0$, $0<\tau<1$, $s=s_C$ and $r=r_0$, then the linearization possesses a double zero eigenvalue, and in this case the bifurcation theorem used in this paper does not apply [15]. We will treat the Hopf bifurcation case and the double eigenvalue case in Parts II and III.

Appendix. In this appendix we regard $L(r)$ as a densely defined, unbounded operator in a Hilbert space H. By making use of the fact that $L(r)$ is a perturbation of a negative-definite, self-adjoint operator A, we can easily deduce some properties of $L(r)$ which will be useful in later work on the bifurcation of periodic solutions, as well as for proving Lemma 1.

The self-adjoint operator A. Let $\lambda>0$, $\mathbf{f}\in J$ and consider the Stokes problem: find $\mathbf{u}\in J$ with $\Delta u\in H^0(\Omega,\mathbb{R}^n)$ such that

$$\text{(A.1)}\qquad \begin{aligned} &-\Pi\Delta\mathbf{u}+\lambda\mathbf{u}=\mathbf{f},\\ &\left.\frac{\partial u}{\partial z}\right|_{z=0,1}=\left.\frac{\partial v}{\partial z}\right|_{z=0,1}=0,\\ &\mathbf{u}\text{ admits periodicity }\Omega. \end{aligned}$$

Then (A.1) is equivalent to a "variational" formulation of the problem [17]: find $\mathbf{u}\in V$ such that

$$\text{(A.2)}\qquad a(\mathbf{u},\mathbf{v})+\lambda(\mathbf{u},\mathbf{v})_0=(\mathbf{f},\mathbf{v})\quad\text{for all }\mathbf{v}\in V,$$

where $a(\mathbf{u},\mathbf{v})=\sum_{i=1}^n\sum_{j=1}^n\int_\Omega D_iu_jD_iv_j\,dx$ defines a bounded symmetric bilinear form on $V\times V$. By classical methods [1], [8] it can be shown that (A.1) has a unique solution $\mathbf{u}\in H^2(\Omega,\mathbb{R}^n)\cap V$, and that $\|\mathbf{u}\|_2\leqq C\|f\|_0$ for some constant C.

If we define the subspace

$$\mathscr{D}(-\Pi\Delta)=\left\{\mathbf{u}\in H^2(\Omega,\mathbb{R}^n)\cap V:\ \left.\frac{\partial u}{\partial z}\right|_{z=0,1}=\left.\frac{\partial v}{\partial z}\right|_{z=0,1}=0\right\},$$

then $-\Pi\Delta:\mathscr{D}(-\Pi\Delta)\to J$ is a densely defined, self-adjoint, closed operator in the Hilbert space J. For real $\lambda>0$, the left-hand side of (A.2) defines a coercive bilinear form on $V\times V$, which implies that any real $\lambda<0$ is in the resolvent set of the operator $(-\Pi\Delta)$. The resolvent operator takes values in V, and the injection of V into J is compact; thus, the resolvent operator $(-\Pi\Delta-\lambda I)^{-1}:J\to J$ is compact. By [12, Thm. II.6.29, p.187], we conclude that the spectrum of $(-\Pi\Delta)$ consists entirely of isolated eigenvalues of finite multiplicities, and that $(-\Pi\Delta-\lambda I)^{-1}$ is compact for every λ not an eigenvalue of $(-\Pi\Delta)$.

Similarly, if we consider the Dirichlet problem

$$\text{(A.3)}\qquad -\Delta\theta+\lambda\theta=f,\quad \theta|_{z=0,1}=0,\quad \theta\text{ admits periodicity }\Omega,$$

and define $\mathscr{D}(-\Delta)=H_0^2(\Omega,\mathbb{R})$, then $-\Delta:\mathscr{D}(-\Delta)\to H^0(\Omega,\mathbb{R})$ is a densely defined, self-adjoint closed operator in $H^0(\Omega,\mathbb{R})$, and every real $\lambda\leqq 0$ is in the resolvent set of $(-\Delta)$, with $(-\Delta-\lambda I)^{-1}$ compact.

Now let $Y=J\times H^0(\Omega,\mathbb{R})\times H^0(\Omega,\mathbb{R})$ and define the operator A by

$$\mathscr{D}(A)=\left\{u=(\mathbf{u},\theta,\zeta)\in H^2(\Omega,\mathbb{R}^{n+2})\cap W:\ \left.\frac{\partial u}{\partial z}\right|_{z=0,1}=\left.\frac{\partial v}{\partial z}\right|_{z=0,1}=0\right\},$$

$$Au=(\sigma\Pi\Delta\mathbf{u},\ \Delta\theta,\ \tau\Delta\zeta)\quad \text{for } u\in\mathscr{D}(A).$$

Then A is a densely defined, closed operator in Y whose spectrum consists entirely of isolated eigenvalues of finite multiplicities; each real $\lambda>0$ is in the resolvent set of A, and $(A-\lambda I)^{-1}: Y\to Y$ is compact for every λ not an eigenvalue of A. In addition, A is self-adjoint, and hence its spectrum $\Sigma(A)$ is a subset of the negative real line, and for every λ not an eigenvalue we have $\|(A-\lambda I)^{-1}\|=1/\operatorname{dist}(\lambda,\Sigma(A))$ [12, p. 272]. By substitution of the Fourier series of §3, it follows that the problem $Au=0$ admits only the trivial solution, and hence $\lambda=0$ is not an eigenvalue of A. Thus the resolvent set of A contains all real $\lambda\geqq 0$. We summarize the properties of A in the following proposition.

PROPOSITION A.1. *The spectrum of A consists entirely of isolated, real, negative eigenvalues of finite multiplicities. For every λ in the resolvent set of A, the resolvent operator $(A-\lambda I)^{-1}$ is compact, and for all real $\lambda>0$, we have the estimate*

$$\|(A-\lambda I)^{-1}\|\leqq\frac{1}{\lambda}. \tag{A.4}$$

The operator $L(r)$. Define the operator $B(r)$ for $r>0$ by

$$\mathscr{D}(B(r))=\mathscr{D}(A),$$

$$B(r)u=(\sigma\Pi[(r\theta-s\zeta)\mathbf{e}],w,w)\quad \text{for } u\in\mathscr{D}(B(r)).$$

Then $B(r)$ is a bounded linear operator in Y and, in particular, $B(r)$ is relatively bounded with respect to A [12, p. 190]: $\mathscr{D}(A)\subset\mathscr{D}(B(r))$ and there exist nonnegative constants a and b such that

$$\|B(r)u\|_0\leqq a\|u\|_0+b\|Au\|_0\quad \text{for all } u\in\mathscr{D}(A). \tag{A.5}$$

In our case we may take $a=\|B(r)\|$ and $b=0$.

Finally, we define $L(r)$ for $r>0$ by

$$\mathscr{D}(L(r))=\mathscr{D}(A),\qquad L(r)=A+B(r).$$

Then $L(r)$ is a densely defined, closed operator in Y. By (A.4), we can choose a real $\lambda>0$ sufficiently large so that

$$a\|(A-\lambda I)^{-1}\|+b<1,$$

where a, b are as in (A.5), and hence by Proposition A.1 and [12, Thm. IV.1.16, p. 196], $(L(r)-\lambda I)^{-1}$ is compact. Applying [12, Thm. III.6.29] again, we obtain the next proposition.

PROPOSITION A.2. *For each $r>0$, the spectrum of $L(r)$ consists entirely of isolated eigenvalues of finite multiplicities. Furthermore, $(L(r)-\lambda I)^{-1}$ is compact for every λ in the resolvent set of $L(r)$.*

Proof of Lemma 1. We can now use Proposition A.2 to prove Lemma 1 of §3. Let $f\in Y$ and consider the solvability of

$$L(r_0)u=f \tag{A.6}$$

for $u \in \mathscr{D}(L(r_0))$ $(=X)$. If λ is in the resolvent set of $L(r_0)$, then (A.6) is equivalent to the problem of finding $u \in Y$ such that

$$u+\lambda(L(r_0)-\lambda I)^{-1}u=(L(r_0)-\lambda I)^{-1}f. \tag{A.7}$$

By Proposition A.2, $(L(r_0)-\lambda I)^{-1}$ is compact for some real $\lambda \neq 0$, and hence by Riesz–Schauder theory a solution exists if and only if

$$\left((L(r_0)-\lambda I)^{-1}f, u^*\right)_0=0$$

for all u^* satisfying $u^*+\lambda[(L(r_0)-\lambda I)^{-1}]^*u^*=0$. But $\lambda \neq 0$ and hence this solvability condition is equivalent to the condition that

$$(f, u^*)_0=0$$

for all u^* satisfying $L(r_0)^*u^*=0$. Thus $R(L(r_0))=N(L(r_0)^*)^{\perp}$.

Acknowledgments. The authors thank the reviewers for many useful suggestions, and Professor Martin Golubitsky for bringing reference [20] to our attention and providing helpful comments.

REFERENCES

[1] J.-P. AUBIN, *Approximation of Elliptic Boundary-Value Problems*, Wiley-Interscience, New York, 1972.
[2] P. G. BAINES AND A. E. GILL, *On thermohaline convection with linear gradients*, J. Fluid Mech., 37 (1969), pp. 289–306.
[3] E. BUZANO AND M. GOLUBITSKY, *Bifurcation on the hexagonal lattice and the planar Bénard problem*, Phil. Trans. Roy. Soc. London, A308 (1983), pp. 617–667.
[4] M. G. CRANDALL AND P. H. RABINOWITZ, *Bifurcation from simple eigenvalues*, J. Functional Anal., 8 (1971), pp. 327–340.
[5] ———, *Bifurcation, perturbation of simple eigenvalues, and linearized stability*, Arch. Rat. Mech. Anal., 52 (1973), pp. 161–180.
[6] L. N. DA COSTA, E. KNOBLOCH AND N. O. WEISS, *Oscillations in double-diffusive convection*, J. Fluid Mech., 109 (1981), pp. 25–43.
[7] H. E. HUPPERT AND D. R. MOORE, *Nonlinear double-diffusive convection*, J. Fluid Mech., 78 (1976), pp. 821–854.
[8] G. IOOSS, *Théorie non linéaire de la stabilité des écoulements laminaires dans le cas de l'échange des stabilités*, Arch. Rat. Mech. Anal., 40 (1971), pp. 166–208.
[9] V. I. IUDOVICH, *On the origin of convection*, Prikl. Mat. Meh., 30 (1966), pp. 1000–1005; J. Appl. Math. Mech., 30 (1966), pp. 1193–1199.
[10] D. D. JOSEPH, *Stability of Fluid Motions* I *and* II, Springer-Verlag, Berlin, 1976.
[11] S. KARLIN, *Total Positivity*, *Vol.* 1, Stanford Univ. Press, Stanford, PA 1968.
[12] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin, 1980.
[13] K. KIRCHGÄSSNER, *Bifurcation in nonlinear hydrodynamic stability*, SIAM Rev., 17 (1975), pp. 652–683.
[14] K. KIRCHGÄSSNER AND H. KIELHÖFER, *Stability and bifurcation in fluid dynamics*, Rocky Mountain J. Math., 3 (1972), pp. 275–318.
[15] E. KNOBLOCH AND M. R. E. PROCTOR, *Nonlinear periodic convection in double-diffusive systems*, J. Fluid Mech., 108 (1981), pp. 291–316.
[16] J. NEU, *Convective flow with subcritical instability*, Phys. Fluids, 25 (1982), pp. 8–13.
[17] P. H. RABINOWITZ, *Existence and nonuniqueness of rectangular solutions of the Bénard problem*, Arch. Rat. Mech. Anal., 29 (1968), pp. 32–57.
[18] R. TEMAM, *Navier–Stokes Equations*, North-Holland, Amsterdam, 1979.
[19] G. VERONIS, *On finite amplitude instability in thermohaline convection*, J. Marine Res., 23 (1965), pp. 1–17.
[20] M. GOLUBITSKY, J. W. SWIFT AND E. KNOBLOCH, *Symmetries and pattern selection in Rayleigh–Bénard convection*, Physica D, to appear (1984).
[21] A. SCHLÜTER, D. LORTZ AND F. BUSSE, *On the stability of steady finite amplitude convection*, J. Fluid Mech., 23 (1965), pp. 129–144.

SIAM J. MATH. ANAL.
Vol. 17, No. 1, January 1986

012

BIFURCATION IN DOUBLY-DIFFUSIVE SYSTEMS II. TIME PERIODIC SOLUTIONS*

WAYNE NAGATA[†‡] AND JAMES W. THOMAS[†]

Abstract. We study a system of double-diffusive convection equations which describe a layer of fluid heated and salted from below. For suitable parameter values Hopf bifurcations of time periodic solutions occur in roll-like, square and hexagonal convection cell patterns. A version of the center manifold theorem suitable for partial differential equations is used to prove the existence of the bifurcating time periodic solutions, and the stability of these solutions is determined from the Poincaré normal form of the reduced equations on the center manifold.

Key words. doubly-diffusive systems, Hopf bifurcation, stability

1. Introduction. This is the second paper in a series of three concerning bifurcations which occur in double-diffusive convection equations. In the first paper [10] (this issue, pp. 91–113), which we will refer to as Part I, we investigated the existence and stability of bifurcating nontrivial equilibrium (steady) solutions. In the present paper we consider bifurcating time periodic solutions of the equations.

We refer to Part I for a description of the double-diffusive convection equations and the cellullar structures of rolls, rectangles, squares and hexagons. Proceeding as in §2 of Part I, we select one of the cellular structures, and then choose the parameters σ, τ and s so that

$$\begin{aligned} &\sigma>0, \qquad 0<\tau<1, \\ &s>\frac{27}{4}\pi^4\tau^2(1+\sigma^{-1})(1-\tau)^{-1}\equiv s_C. \end{aligned} \tag{1.1}$$

Then when

$$r=\left(\frac{\sigma+\tau}{\sigma+1}\right)s+\frac{27}{4}\pi^4(1+\tau)(1+\tau\sigma^{-1})\equiv r_H, \tag{1.2}$$

the linearization of the convection equations about the constant gradient solution possesses two simple eigenvalues $\lambda=\pm i\omega_0$, where

$$\omega_0^2=\frac{9}{4}\pi^4(\sigma+\tau+\sigma\tau)-\frac{1}{3}\sigma(r_H-s). \tag{1.3}$$

All the other eigenvalues of the linearization have negative real parts when $r=r_H$ [5]. We expect a Hopf bifurcation of time periodic solutions to occur at $r=r_H$ in the nonlinear convection equations.

The existence of Hopf bifurcations in *Navier–Stokes* type systems can be proven within a number of theoretical settings, for example [1], [3], [4], [6], [7], [8], [9]. These settings also provide means for computing the stability of the bifurcating solutions near the point of bifurcation. In this paper we use center manifold methods (see [3], [4], and [9]), which we feel are conceptually more natural.

In §2 we formulate the nonlinear convection equations as an abstract evolution equation in a Hilbert space, and prove the existence of bifurcating time periodic

* Received by the editors January 16, 1984, and in revised form July 10, 1984.
† Department of Mathematics, Colorado State University, Fort Collins, Colorado 80523.
‡ Present address: Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912.

solutions for each cellular structure. Then in §3 we compute for rolls, squares and hexagons the direction of bifurcation and the stability of the bifurcating solutions with respect to small perturbations having the same cellular structure. Our results for rolls agree with the results of formal computations performed previously by [5] and [2]. A simplified model corresponding to rolls with different dimensions was treated by [11].

2. Hopf bifurcations of time periodic solutions. In this section we express the double-diffusive convection equations as a one-parameter family of semilinear parabolic equations in a Hilbert space. The family satisfies hypotheses of a version of the center manifold theorem [4], and from this theorem follows the existence of Hopf bifurcations.

In §3 and the appendix of Part I we defined the Hilbert spaces X and Y and the operators A, $B(r)$ and M. We continue with the same notation. Since the unbounded, densely defined operator A is self-adjoint and negative definite, it follows that the fractional powers $(-A)^\alpha$ and the Banach spaces $X^\alpha = \mathscr{D}((-A)^\alpha)$ with norms $\|u\|_\alpha = \|(-A)^\alpha u\|_\gamma$ are well defined for $\alpha \geqq 0$ [4, p. 19]. If $\alpha = 0$, then $X^0 = Y$, and if $\alpha = 1$, then $X^1 = X = \mathscr{D}(A)$. Furthermore, if $0 \leqq \alpha < 1$, then $X \subset X^\alpha \subset Y$ with continuous imbeddings. Thus, the bounded linear operator $B(r)$ in Y defines an analytic mapping $(u,r) \mapsto B(r)u$ from $X^\alpha \times \mathbb{R}$ into Y, when $0 \leqq \alpha < 1$. The nonlinear operator M is a bounded bilinear operator from $X^\alpha \times X^\alpha$ into Y, when $\alpha > \frac{3}{4}$ [4, p. 79]. Thus,

$$f(u,r) = B(r)u + M(u,u) \tag{2.1}$$

defines an analytic operator $f\colon X^\alpha \times \mathbb{R} \to Y$. We observe that $f(0,r) = 0$ for all $r \in \mathbb{R}$, and $D_u f(0,r) = B(r)$.

We now write the double-diffusive convection equations in the abstract form

$$\frac{du}{dt} = Au + f(u,r) \tag{2.2}$$

for $u \in Y$, where $(-A)$ is a sectorial operator in Y and f is analytic when $\alpha > \frac{3}{4}$. Equation (2.2) generates a unique parametrized family of local semiflows in X^α for $\frac{3}{4} < \alpha < 1$ [4, p. 54], which is jointly analytic in the initial condition $u(0) \in X^\alpha$, the parameter $r \in \mathbb{R}$ and time $t > 0$ [4, p. 66].

When (1.1) and (1.2) are satisfied, the linearization $L(r_H) \equiv A + D_u f(0, r_H)$ at $r = r_H$ possesses two simple eigenvalues $\pm i\omega_0$ where ω_0 is given by (1.3), and all other eigenvalues of $L(r_H)$ have negative real parts as discussed above. Since the spectrum of $L(r)$ for $r \in \mathbb{R}$ is a closed subset of the complex plane consisting entirely of isolated eigenvalues which depend analytically on r, near $r = r_H$ the spectrum of $L(r)$ consists of

$$\{\alpha(r) \pm i\omega(r)\} \cup \Sigma',$$

where $\Sigma' \subset \{\operatorname{Re}\lambda < \beta < 0\}$, $\alpha(r_H) = 0$ and $\omega(r_H) = \omega_0$. By implicitly differentiating the dispersion relation for λ [Part I, (3.8)] with resepct to r, we obtain the nondegeneracy condtion

$$\alpha'(r_H) > 0, \tag{2.3}$$

i.e., the critical eigenvalues cross into the positive complex half plane with nonzero speed. It then follows from [4, p. 181] that (2.2) has a two-dimensional local invariant manifold S_r containing $0 \in X^\alpha$ for r near r_H. Local bifurcation and stability of solutions of (2.2) is determined by local bifurcation and stability of solutions for the flow in S_r.

Near the origin, the flow in S_r for r near r_H is described by

$$\begin{pmatrix} \dot{y}_1 \\ \dot{y}_2 \end{pmatrix} = \begin{bmatrix} \alpha(r) & \omega(r) \\ -\omega(r) & \alpha(r) \end{bmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} + \begin{pmatrix} \phi_1(y_1, y_2; r) \\ \phi_2(y_1, y_2; r) \end{pmatrix} \tag{2.4}$$

where $\alpha(r_H)=0$, $\alpha'(r_H)>0$, $\omega(r_H)=\omega_0>0$ and $\phi_{1,2}(y_1,y_2;r)=O(y_1^2+y_2^2)$ as $(y_1,y_2)\to(0,0)$. Under the additional assumption

(A) the origin is not a center for (2.4) when $r=r_H$, and this is true independent of terms of order $|y_1|^p+|y_2|^p$, for some $p\geqq 2$, Henry then concludes that (2.2) either has a supercritical or subcritical branch of time periodic solutions near $r=r_H$. More precisely, we have the following result.

THEOREM 1. *Let Ω be one of the following fundamental domains of spatial periodicity*:

i) (*Rolls*) $\Omega=C_1\times(0;1)\subset\mathbb{R}^2$, *where* $C_1=(0,2\sqrt{2})$,

ii) (*Rectangles*) $\Omega=C_2\times(0,1)\subset\mathbb{R}^3$, *where* $C_2=(0,2\pi/\alpha_1)\times(0,2\pi/\alpha_2)$ *and* $\alpha_1^2+\alpha_2^2=\pi^2/2$, $\alpha_1,\alpha_2>0$, $\alpha_2\neq\alpha_1\sqrt{m^2-1}$ *for all* $m\in N$,

iii) (*Squares*) $\Omega=C_3\times(0,1)\subset\mathbb{R}^3$, *where* $C_3=(0,4)\times(0,4)$,

iv) (*Hexagons*) $\Omega=C_4\times(0,1)\subset\mathbb{R}^3$, *where* C_4 *is the interior of a regular hexagon with sides of length* $4\sqrt{2}/3$.

Then for each fixed $\sigma\in(0,\infty)$, $\tau\in(0,1)$ *and* $s\in(s_c,\infty)$, (2.6) *can be reduced to the one-parameter family of equations* (2.4) *on local invariant manifolds* S_r *for* r *in some neighborhood of* r_H. *Moreover, if* (A) *is satisfied then one of the following two conclusions must hold* [4, *p.* 182]:

a. (*Supercritical Hopf bifurcation*) *The origin is asymptotically stable for* (2.2) *when* $r=r_H$, *and there is an* $\varepsilon_0>0$ *such that there is no periodic orbit of* (2.2) *in a neighborhood of the origin in* X^α *for* $r\in[r_H-\varepsilon_0,r_H]$, *but a unique family of orbitally asymptotically stable periodic orbits grows out of the origin for* $r\in(r_H,r_H+\varepsilon_0]$; *or*

b. (*Supercritical Hopf bifurcation*) *The origin is unstable for* (2.2) *when* $r=r_H$, *and there is an* $\varepsilon_0>0$ *such that for* $r\in[r_H-\varepsilon_0,r_H)$ *a unique family of unstable periodic orbits of* (2.2) *shrinks to the origin as* $r\to r_H-$, *but there is no periodic orbit of* (2.2) *in a neighborhood of the origin for* $r\in[r_H,r_H+\varepsilon_0]$.

In both cases the periods of the nontrivial periodic orbits approach $2\pi/\omega_0$ *as* $r\to r_H$.

In the next section we will show that there exist $\sigma\in(0,\infty)$, $\tau\in(0,1)$ and $s\in(s_c,\infty)$ such that assumption (A) is satisfied. In addition, we will be able to determine whether we have supercritical or subcritical Hopf bifurcations.

3. Stability of the bifurcating time periodic solutions. In this section we determine whether the bifurcating time periodic solutions described in the last section are orbitally asymptotically stable or are unstable. Following Hassard, Kazarinoff and Wan [3] we express the family of equation (2.4) in terms of the complex variable $z=y_1+iy_2$ as

$$\dot{z}=\lambda(r)z+g(z,\bar{z};r), \tag{3.1}$$

where $\bar{z}=y_1-iy_2$, $\lambda(r)=\alpha(r)+i\omega(r)$, $\lambda(r_H)=i\omega_0$, $\alpha'(r_H)>0$ and $g(z,\bar{z};r)=0(|z|^2)$ as $z\to 0$. By transforming (3.1) into the Poincaré normal form

$$\dot{\xi}=\lambda(r)\xi+c_1(r)\xi|\xi|^2+O\left(|\xi|\,|(\xi,r-r_H)|^4\right), \tag{3.2}$$

one can determine whether the Hopf bifurcation is supercritical or subcritical. If $\operatorname{Re}c_1(r_H)\neq 0$ then the origin is not a center for (3.2), and hence the origin is not a center for (3.1), when $r=r_H$. Furthermore, this is true independent of terms of order $|z|^4$

in (3.1). If we write

$$g(z,\bar{z};r)=\sum_{j+k=2}^{3} g_{jk}(r)\frac{z^j\bar{z}^k}{j!k!}+O\left(|z|^4\right) \tag{3.3}$$

for the term appearing in (3.1), then we can use the formula [3, p. 47]

$$c_1(r_H)=\frac{i}{2\omega_0}\left[g_{20}(r_H)g_{11}(r_H)-2|g_{11}(r_H)|^2-\frac{1}{3}|g_{02}(r_H)|^2\right]+\frac{1}{2}g_{21}(r_H). \tag{3.4}$$

The Hopf bifurcation is supercritical or subcritical according as $(\mathrm{Re}\,c_1(r_H))/\alpha'(r_H)$ is negative or positive, and the bifurcating periodic solutions for r near r_H are orbitally asymptotically stable with asymptotic phase or unstable according as $\mathrm{Re}\,c_1(r_H)$ is negative or positive. Thus to determine the stabilty of the bifurcating periodic orbits we must find the leading terms of the nonlinear part (3.3) of the equation on the center manifold (3.1).

To find $c_1(r_H)$, it suffices to reduce (2.2) to the center manifold at $r=r_H$. We follow the procedure of Hassard, Kazarinoff and Wan [3, Chap. 5], which we now outline. First, we rewrite (2.2) as

$$\frac{du}{dt}=L(r)u+M(u,u), \tag{3.5}$$

where $L(r)=A+D_uf(0,r)$. Next, we solve the eigenvalue problems

$$L(r_H)q=i\omega_0 q \tag{3.6}$$

and

$$L(r_h)^*q^*=-i\omega_0 q^*, \tag{3.7}$$

where the adjoint $L(r)^*$ of $L(r)$ is given explicitly in Part I.

We can then write X^α as the direct sum $X^\alpha=X_c\oplus X_s^\alpha$, where

$$X_c=\{zq+\bar{z}\bar{q}:\ z\in C\},\qquad X_s^\alpha=\{w\in X^\alpha:\ (w,q^*)=0\} \tag{3.8}$$

and $(\cdot,\cdot)$ is the restriction of the Y-inner product to X^α. Observe that this restriction is continuous with respect to the norm $\|\cdot\|_\alpha$ when $\alpha\geqq 0$, due to the continuity of the injection $X^\alpha\subset Y$. Thus for any $u\in X^\alpha$ we have

$$u=zq+\bar{z}\bar{q}+w \tag{3.9}$$

where $z=(u,q^*)$ and $w=u-(u,q^*)q-(u,\bar{q}^*)\bar{q}$ provided we have normalized $(q,q^*)=1$. Projecting (3.6) onto $\mathbb{C}\cong X_c^\alpha$ and X_s^α, we obtain

$$\frac{dz}{dt}=i\omega_0 z+(M(u,u),q^*) \tag{3.10}$$

and

$$\frac{dw}{dt}=L(r_H)w+H(z,\bar{z},w) \tag{3.11}$$

respectively, where u is given by (3.9) and

$$H(z,\bar{z},w)=M(u,u)=(M(u,u),q^*)q-(M(u,u),\bar{q}^*)\bar{q}. \tag{3.12}$$

The center manifold S_{r_H} can be expressed locally in terms of a function $w(z,\bar{z};r_H) \in X_s^\alpha$,

$$S_{r_H} = \{ zq + \bar{z}\bar{q} + w(z,\bar{z};r_H) : |z| < \delta \} \tag{3.13}$$

where $w(0,0;r_H)=0$ and $Dw(0,0;r_H)=0$. Since $f(u,r_H)$ is analytic in $u \in X^\alpha$, $w(z,\bar{z};r_H)$ can be approximated to arbitrarily high order p by a polynomial [4, p. 171]

$$w(z,\bar{z};r_H) = \sum_{j+k=2}^{p} \frac{z^j \bar{z}^k}{j!k!} w_{jk} + O(|z|^p). \tag{3.14}$$

Equation (3.1) with $r=r_H$ on the center manifold S_{r_H} is obtained by setting $u = zq + \bar{z}\bar{q} + w(z,\bar{z};r_H)$ in (3.10). Thus locally we have

$$g(z,z;r_H) = (M(zq+\bar{z}\bar{q}+w(z,z;r_H), zq+\bar{z}\bar{q}+w(z,z;r_H)), q^*). \tag{3.15}$$

To lowest order $u = zq + \bar{z}\bar{q} + O(|z|^2)$, and hence

$$M(u,u) = z^2 M(q,q) + 2z\bar{z}\,\mathrm{Re}\,M(q,\bar{q}) + \bar{z}^2 M(\bar{q},\bar{q}) + O(|z|^3). \tag{3.16}$$

Due to the boundary conditions, the eigenvalue problems (3.6) and (3.7) can be solved exactly for q and q^*, and then a calculation will show that $M(q,q)$, $\mathrm{Re}\,M(q,\bar{q})$ and $M(\bar{q},\bar{q})$ are all orthogonal to q^* (see the example for rolls later in this section). Hence

$$(M(u,u),q^*) = O(|z|^3), \tag{3.17}$$

$$g_{20}(r_H) = g_{11}(r_H) = g_{02}(r_H) = 0 \tag{3.18}$$

and

$$H(z,\bar{z},w(z,\bar{z};r_H)) = M(u,u) + O(|z|^3), \tag{3.19}$$

where $M(u,u)$ is given by (3.16). To determine the stability of the bifurcating solutions it follows from (3.5) and (3.18) that we only need to find $g_{21}(r_H)$. To do this we need the quadratic approximation to $w(z,\bar{z};r_H)$,

$$w(z,z;r_H) = \frac{1}{2} z^2 w_{20} + z\bar{z} w_{11} + \frac{1}{2} \bar{z}^2 w_{02} + O(|z|^3). \tag{3.20}$$

To find w_{20}, w_{11} and w_{02} we solve [3, p. 237]

$$\begin{aligned} (L(r_H) - 2i\omega_0) w_{20} &= -2M(q,q), \\ L(r_H) w_{11} &= -2\,\mathrm{Re}\,M(q,\bar{q}), \end{aligned} \tag{3.21}$$

and $w_{02} = \bar{w}_{20}$. Substituting (3.20) into (3.15) and collecting the coefficients in $\bar{z}^2 z$, we obtain

$$g_{21}(r_H) = (M(w_{20},\bar{q}) + M(\bar{q},w_{20}), q^*) + 2(M(w_{11},q) + M(q,w_{11}), q^*). \tag{3.22}$$

Since $\alpha'(r_H) > 0$ for $\sigma > 0$, $0 < \tau < 1$ and $s > s_c$, the Hopf bifurcation is supercritical or subcritical according to whether $\mathrm{Re}\,g_{21}(r_H)$ is negative or positive, and we have the following theorem.

THEOREM 2. *If* $\mathrm{Re}\,g_{21}(r_H) \neq 0$, *then assumption* (A) *of* §2 *is satisfied. Furthermore, if* $\mathrm{Re}\,g_{21}(r_H) < 0$ *then conclusion* (a) *of Theorem* 1 *holds, while if* $\mathrm{Re}\,g_{21}(r_H) > 0$ *then conclusion* (b) *holds.*

If $\mathrm{Re}\,g_{21}(r_H) = 0$ then higher order terms in (5.2) are required to see if (A) is satisfied.

We now illustrate the procedure of finding an explicit expression for $g_{21}(r_H)$ in the case of rolls. For $\Omega = C_1 \times (0,1)$ in $\mathbb{R}^2$ the components of an element $q \in Y$ have Fourier series expansions (see Part I)

$$
\begin{aligned}
q_1(x_1,x_3) &= \sum_{n_1=1}^{\infty} q_1(n_1,x_3)\sin n_1\alpha_1 x_1,\\
q_3(x_1,x_3) &= \sum_{n_1=0}^{\infty} q_3(n_1,x_3)\cos n_1\alpha_1 x_1,\\
q_4(x_1,x_3) &= \sum_{n_1=0}^{\infty} q_4(n_1,x_3)\cos n_1\alpha_1 x_1,\\
q_5(x_1,x_3) &= \sum_{n_1=0}^{\infty} q_5(n_1,x_3)\cos n_1\alpha_1 x_1,
\end{aligned}
\tag{3.23}
$$

where $\alpha_1 = \pi/\sqrt{2}$. Then the eigenvalue problems (3.6) and (3.7) become

(3.24)

$$
\begin{aligned}
&\sigma\Delta(n_1)q_1(n_1,x_3)+\sigma n_1\alpha_1 p(n_1,x_3) = i\omega_0 q_1(n_1,x_3),\\
&\sigma\Delta(n_1)q_3(n_1,x_3)-\sigma\frac{d}{dx_3}p(n_1,x_3)+\sigma r_H q_4(n_1,x_3)-\sigma s q_5(n_1,x_3) = i\omega_0 q_3(n_1,x_3),\\
&\Delta(n_1)q_4(n_1,x_3)+q_3(n_1,x_3) = i\omega_0 q_4(n_1,x_3),\\
&\tau\Delta(n_1)q_5(n_1,x_3)+q_3(n_1,x_3) = i\omega_0 q_5(n_1,x_3),\\
&n_1\alpha_1 q_1(n_1,x_3)+\frac{d}{dx_3}q_3(n_1,x_3) = 0 \quad \text{for } x_3 \in (0,1),\\
&\frac{d}{dx_3}q_1(n_1,x_3)\Big|_{x_3=0,1} = q_3(n_1,x_3)|_{x_3=0,1} = q_4(n_1,x_3)|_{x_3=0,1} = q_5(n_1,x_3)|_{x_3=0,1} = 0,
\end{aligned}
$$

and

(3.25)

$$
\begin{aligned}
&\sigma\Delta(n_1)q^*{}_1(n_1,x_3)+\sigma n_1\alpha_1 p^*(n_1,x_3) = -i\omega_0 q^*{}_1(n_1,x_3),\\
&\sigma\Delta(n_1)q^*{}_3(n_1,x_3)-\sigma\frac{d}{dx_3}p^*(n_1,x_3)+q^*{}_4(n_1,x_3)+q^*{}_5(n_1,x_3) = -i\omega_0 q^*{}_3(n_1,x_3),\\
&\Delta(n_1)q^*{}_4(n_1,x_3)+\sigma r_H q^*{}_3(n_1,x_3) = -i\omega_0 q^*{}_4(n_1,x_3),\\
&\tau\Delta(n_1)q^*{}_5(n_1,x_3)-\sigma s q^*{}_3(n_1,x_3) = -i\omega_0 q^*{}_5(n_1,x_3),\\
&n_1\alpha_1 q^*{}_1(n_1,x_3)+\frac{d}{dx_3}q^*{}_3(n_1,x_3) = 0 \quad \text{for } x_3 \in (0,1),\\
&\frac{d}{dx_3}q^*{}_1(n_1,x_3)\Big|_{x_3=0,1} = q^*{}_3(n_1,x_3)|_{x_3=0,1} = q^*{}_4(n_1,x_3)|_{x_3=0,1} = q^*{}_5(n_1,x_3)|_{x_3=0,1} = 0,
\end{aligned}
$$

for $n_1 = 0,1,2,\cdots$, where $\Delta(n_1) = (d/dx_3) - n_1^2\alpha_1^2$ and p, p^* are fluid pressures. The pressures can be eliminated, and the equations (3.24), (3.25) can be reduced to single equations for $q_3(n_1,x_3)$, $q^*{}_3(n_1,x_3)$ as in Part I. The boundary conditions allow the

Fourier series expansions

$$
\begin{aligned}
q_3(n_1,x_3) &= \sum_{n_3=1}^{\infty} q_3(n_1,n_3)\sin n_3\pi x_3,\\
q^*{}_3(n_1,x_3) &= \sum_{n_3=1}^{\infty} q^*{}_3(n_1,x_3)\sin n_3\pi x_3,
\end{aligned}
\tag{3.26}
$$

for $n_1 = 0, 1, 2, \cdots$. Equations (3.24), (3.25) have no nontrivial solutions unless $n_1 = n_3 = 1$ and we obtain the eigenfunctions in the case of rolls,

$$
q = \begin{pmatrix}
-\dfrac{2\alpha_1}{\pi}\sin\alpha_1 x_1\cos\pi x_3\\
\cos\alpha_1 x_1\sin\pi x_3\\
\left(\dfrac{3\pi^2}{2}+i\omega_0\right)^{-1}\cos\alpha_1 x_1\sin\pi x_3\\
\left(\dfrac{3\pi^2\tau}{2}+i\omega_0\right)^{-1}\cos\alpha_1 x_1\sin\pi x_3
\end{pmatrix},
\tag{3.27}
$$

$$
q^* = (\bar{N})^{-1}\begin{pmatrix}
-\dfrac{2\alpha_1}{\pi}\sin\alpha_1 x_1\cos\pi x_3\\
\cos\alpha_1 x_1\sin\pi x_3\\
\sigma r_H\left(\dfrac{3\pi^2}{2}-i\omega_0\right)^{-1}\cos\alpha_1 x_1\sin\pi x_3\\
-\sigma s\left(\dfrac{3\pi\tau}{2}-i\omega_0\right)^{-1}\cos\alpha_1 x_1\sin\pi x_3
\end{pmatrix}
\tag{3.28}
$$

where

$$
N = \frac{1}{2}\sqrt{2}\left[3+\sigma r_H\left(\frac{3\pi^2}{2}+i\omega_0\right)^{-2}-\sigma s\left(\frac{3\pi^2}{2}+i\omega_0\right)^{-2}\right]
\tag{3.29}
$$

is chosen so that

$$
(q,q^*) = \int_0^1\int_0^{\alpha_1/2\pi}(q_1\bar{q}^*{}_1+q_3\bar{q}^*{}_3+q_4\bar{q}^*{}_4+q_5\bar{q}^*{}_5)\,dx_1\,dx_3 = 1.
$$

We then compute

$$
(\nabla\cdot\underset{\sim}{q})\,q = \begin{pmatrix}
\alpha_1\sin 2\alpha_1 x_1\\
\dfrac{\pi}{2}\sin 2\pi x_3\\
\dfrac{\pi}{2}\left(\dfrac{3\pi^2}{2}+i\omega_0\right)^{-1}\sin 2\pi x_3\\
\dfrac{\pi}{2}\left(\dfrac{3\pi^2\tau}{2}+i\omega_0\right)^{-1}\sin 2\pi x_3
\end{pmatrix}
\tag{3.30}
$$

and

$$(\nabla\cdot\underset{\sim}{q})\bar{q}=\begin{pmatrix}\alpha_1\sin 2\alpha_1 x_1\\ \frac{\pi}{2}\sin 2\pi x_3\\ \frac{\pi}{2}\left(\frac{3\pi^2}{2}-i\omega_0\right)^{-1}\sin 2\pi x_3\\ \frac{\pi}{2}\left(\frac{3\pi^2\tau}{2}-i\omega_0\right)^{-1}\sin 2\pi x_3\end{pmatrix}. \tag{3.31}$$

Note that $(M(q,q),q)=(-\Pi_0(\nabla\cdot q)q,q)=(-(\nabla\cdot q)q,q)=0$, and similarly $(M(q,\bar{q}),q^*)=0$ since $\int_0^1\cos\pi x_3\,dx_3=0$ and $\int_0^1\sin\pi x_3\sin 2\pi x_3\,dx_3=0$. This verifies in the case of rolls the assertion following (3.16). For squares and hexagons, the assertion is true for similar reasons. We then solve (3.21) for w_{20} and w_{11} by finding their Fourier coefficients as we did for q and q^*, and obtain

$$w_2^0=\begin{pmatrix}0\\ 0\\ -(4\pi^2+2i\omega_0)^{-1}\left(\frac{3\pi^2}{2}+i\omega_0\right)^{-1}\pi\sin 2\pi x_3\\ -(4\pi^2\tau+2i\omega_0)^{-1}\left(\frac{3\pi^2\tau}{2}+i\omega_0\right)^{-1}\pi\sin 2\pi x_3\end{pmatrix},$$

$$w_{11}=\begin{pmatrix}0\\ 0\\ -(4\pi^2)^{-1}\mathrm{Re}\left(\frac{3\pi^2}{2}+i\omega_0\right)^{-1}\pi\sin 2\pi x_3\\ -(4\pi^2\tau)^{-1}\mathrm{Re}\left(\frac{3\pi^2\tau}{2}+i\omega_0\right)^{-1}\pi\sin 2\pi x_3\end{pmatrix}. \tag{3.32}$$

Finally from (3.22), (3.27), (3.28) and (3.32) we obtain the expression

(3.33)

$$\begin{aligned}g_{21}(r_H)=\frac{\sqrt{2}\pi}{2N}\Bigg\{&-\sigma r_H\left(\frac{3\pi^2}{2}+i\omega_0\right)^{-1}\Bigg[(4\pi^2+2i\omega_0)^{-1}\left(\frac{3\pi^2}{2}+i\omega_0\right)^{-1}\\ &+2(4\pi^2)^{-1}\mathrm{Re}\left(\frac{3\pi^2}{2}+i\omega_0\right)^{-1}\Bigg]\\ &+\sigma s\left(\frac{3\pi^2}{2}+i\omega_0\right)^{-1}\Bigg[(4\pi^2+2i\omega_0)^{-1}\left(\frac{3\pi^2\tau}{2}+i\omega_0\right)^{-1}\\ &+2(4\pi^2)^{-1}\mathrm{Re}\left(\frac{3\pi^2\tau}{2}+i\omega_0\right)^{-1}\Bigg]\Bigg\}.\end{aligned}$$

The computations to find $g_{21}(r_H)$ in the cases of squares and hexagons proceed similarly, but are somewhat longer. Here we state only the results: for squares we have

(3.34)

$$
\begin{aligned}
g_{21}(r_H) = \frac{2\pi^2}{N}\Bigg\langle & 3(A+2B) \\
& - \sigma r_H\left(\frac{3\pi^2}{2}+i\omega_0\right)^{-1}\Bigg\{4\left(4\pi^2+2i\omega_0\right)^{-1}\left(\frac{3\pi^2}{2}+i\omega_0\right)^{-1} \\
& \qquad + 8(4\pi^2)^{-1}\,\mathrm{Re}\left(\frac{3\pi^2}{2}+i\omega_0\right)^{-1} \\
& \qquad + \left(5\pi^2+2i\omega_0\right)^{-1}\left[\left(\frac{3\pi^2}{2}+i\omega_0\right)^{-1}-A\right] \\
& \qquad + 2(5\pi^2)^{-1}\left[\mathrm{Re}\left(\frac{3\pi^2}{2}+i\omega_0\right)^{-1}-B\right]\Bigg\} \\
& + \sigma s\left(\frac{3\pi^2\tau}{2}+i\omega_0\right)^{-1}\Bigg\{4\left(4\pi^2\tau+2i\omega_0\right)^{-1}\left(\frac{3\pi^2\tau}{2}+i\omega_0\right)^{-1} \\
& \qquad + 8(4\pi^2\tau)^{-1}\,\mathrm{Re}\left(\frac{3\pi^2\tau}{2}+i\omega_0\right)^{-1} \\
& \qquad + \left(5\pi^2\tau+2i\omega_0\right)^{-1}\left[\left(\frac{3\pi^2\tau}{2}+i\omega_0\right)^{-1}-A\right] \\
& \qquad + 2(5\pi^2\tau)-1\left[\mathrm{Re}\left(\frac{3\pi^2\tau}{2}+i\omega_0\right)^{-1}-B\right]\Bigg\}\Bigg\rangle .
\end{aligned}
$$

where

$$
N = 3+\sigma r_H\left(\frac{3\pi^2}{2}+i\omega_0\right)^{-2}-\sigma s\left(\frac{3\pi^2\tau}{2}+i\omega_0\right)^{-2},
$$

$$
\begin{aligned}
A = & \left[-\left(25\pi^2\sigma+10i\omega_0\right)+\sigma r_H\left(5\pi^2+2i\omega_0\right)^{-1}-\sigma s\left(5\pi^2\tau+2i\omega_0\right)^{-1}\right]^{-1} \\
& \cdot\left[3+\sigma r_H\left(\frac{3\pi^2}{2}+i\omega_0\right)^{-1}\left(5\pi^2+2i\omega_0\right)^{-1}\right. \\
& \qquad \left.-\sigma s\left(\frac{3\pi^2\tau}{2}+i\omega_0\right)^{-1}\left(5\pi^2\tau+2i\omega_0\right)^{-1}\right],
\end{aligned}
$$

and

$$B=\left[-25\pi^2\sigma+\sigma r_H(5\pi^2)^{-1}-\sigma s(5\pi^2\tau)^{-1}\right]^{-1}$$
$$\cdot\left[3+\sigma r_H(5\pi^2)^{-1}\mathrm{Re}\left(\frac{3\pi^2}{2}+i\omega_0\right)^{-1}-\sigma s(5\pi^2\tau)^{-1}\mathrm{Re}\left(\frac{3\pi^2\tau}{2}+i\omega_0\right)^{-1}\right],$$

while for hexagons

(3.35)

$$g_{21}(r_H)=\frac{\pi^2}{2N}\Bigg\langle 9A'+3A''+18B'+6B''$$
$$-\sigma r_H\left(\frac{3\pi^2}{2}+i\omega_0\right)^{-1}\Bigg\{24(4\pi^2+2i\omega_0)^{-1}\left(\frac{3\pi^2}{2}+i\omega_0\right)^{-1}$$
$$+48(4\pi^2)^{-1}\mathrm{Re}\left(\frac{3\pi^2}{2}+i\omega_0\right)^{-1}$$
$$+3\left(\frac{9\pi^2}{2}+2i\omega_0\right)^{-1}\left[3\left(\frac{3\pi^2}{2}+i\omega_0\right)^{-1}-A'\right]$$
$$+6\left(\frac{9\pi^2}{2}\right)\left[3\,\mathrm{Re}\left(\frac{3\pi^2}{2}+i\omega_0\right)^{-1}-B'\right]$$
$$+\left(\frac{11\pi^2}{2}+2i\omega_0\right)^{-1}\left[\left(\frac{3\pi^2}{2}+i\omega_0\right)^{-1}-A'\right]$$
$$+2\left(\frac{11\pi^2}{2}\right)^{-1}\left[\mathrm{Re}\left(\frac{3\pi^2}{2}+i\omega_0\right)^{-1}-B''\right]\Bigg\}$$
$$+\sigma s\left(\frac{3\pi^2\tau}{2}+i\omega_0\right)^{-1}\Bigg\{24(4\pi^2\tau+2i\omega_0)^{-1}\left(\frac{3\pi^2\tau}{2}+i\omega_0\right)^{-1}$$
$$+48(4\pi^2\tau)^{-1}\mathrm{Re}\left(\frac{3\pi^2\tau}{2}+i\omega_0\right)^{-1}$$
$$+3\left(\frac{9\pi^2\tau}{2}+2i\omega_0\right)^{-1}\left[3\left(\frac{3\pi^2\tau}{2}+i\omega_0\right)^{-1}-A'\right]$$
$$+6\left(\frac{9\pi^2\tau}{2}\right)^{-1}\left[3\,\mathrm{Re}\left(\frac{3\pi^2\tau}{2}+i\omega_0\right)^{-1}-B'\right]$$
$$+\left(\frac{11\pi^2\tau}{2}+2i\omega_0\right)^{-1}\left[\left(\frac{3\pi^2\tau}{2}+i\omega_0\right)^{-1}-A''\right]$$
$$+2\left(\frac{11\pi^2\tau}{2}\right)^{-1}\left[\mathrm{Re}\left(\frac{3\pi^2\tau}{2}+i\omega_0\right)^{-1}-B''\right]\Bigg\}\Bigg\rangle$$

where

$$N = 3 + \sigma r\left(\frac{3\pi^2}{2} + i\omega_0\right)^{-2} - \sigma s\left(\frac{3\pi^2\tau}{2} + i\omega_0\right)^{-2},$$

$$A' = 3\left[-9\left(\frac{9\pi^2\sigma}{2} + 2i\omega_0\right) + \sigma r_H\left(\frac{9\pi^2}{2} + 2i\omega_0\right)^{-1} - \sigma s\left(\frac{9\pi^2\tau}{2} + 2i\omega_0\right)^{-1}\right]^{-1}$$
$$\cdot\left[3 + \sigma r_H\left(\frac{3\pi^2}{2} + i\omega_0\right)^{-1}\left(\frac{9\pi^2}{2} + 2i\omega_0\right)^{-1} - \sigma s\left(\frac{3\pi^2\tau}{2} + i\omega_0\right)^{-1}\left(\frac{9\pi^2\tau}{2} + 2i\omega_0\right)^{-1}\right],$$

$$A'' = \left[-\frac{11}{3}\left(\frac{11\pi^2\sigma}{2} + 2i\omega_0\right) + \sigma r_H\left(\frac{11\pi^2}{2} + 2i\omega_0\right)^{-1} - \sigma s\left(\frac{11\pi^2\tau}{2} + 2i\omega_0\right)^{-1}\right]^{-1}$$
$$\cdot\left[3 + \sigma r_H\left(\frac{3\pi^2}{2} + i\omega_0\right)^{-1}\left(\frac{11\pi^2}{2} + 2i\omega_0\right)^{-1} - \sigma s\left(\frac{3\pi^2\tau}{2} + i\omega_0\right)^{-1}\left(\frac{11\pi^2\tau}{2} + 2i\omega_0\right)^{-1}\right],$$

$$B' = 3\left[-9\left(\frac{9\pi^2\sigma}{2}\right) + \sigma r_H\left(\frac{9\pi^2}{2}\right)^{-1} - \sigma s\left(\frac{9\pi^2\tau}{2}\right)^{-1}\right]^{-1}$$
$$\cdot\left[3 + \sigma r_H\left(\frac{9\pi^2}{2}\right)^{-1}\mathrm{Re}\left(\frac{3\pi^2}{2} + i\omega_0\right)^{-1} - \sigma s\left(\frac{9\pi^2\tau}{2}\right)^{-1}\mathrm{Re}\left(\frac{3\pi^2\tau}{2} + i\omega_0\right)^{-1}\right],$$

$$B'' = \left[-\frac{11}{3}\left(\frac{11\pi^2\sigma}{2}\right) + \sigma r_H\left(\frac{11\pi^2}{2}\right)^{-1} - \sigma s\left(\frac{11\pi^2\tau}{2}\right)^{-1}\right]^{-1}$$
$$\cdot\left[3 + \sigma r_H\left(\frac{11\pi^2}{2}\right)^{-1}\mathrm{Re}\left(\frac{3\pi^2}{2} + i\omega_0\right)^{-1} - \sigma s\left(\frac{11\pi^2\tau}{2}\right)\mathrm{Re}\left(\frac{3\pi^2\tau}{2} + i\omega_0\right)^{-1}\right]^{-1}.$$

The dependence of $\mathrm{Re}\, g_{21}(r_H)$ on σ, τ and s does not appear to be obvious from (3.33), (3.34) or (3.35) so we evaluated $\mathrm{Re}\, g_{21}(r_H)$ numerically. For all four spatially periodic structures—rolls, rectangles, squares and hexagons—the locus of points in $\sigma\tau$ s-space where $\mathrm{Re}\, g_{21}(r_H) = 0$ appears to be a two-dimensional surface in the three-dimensional parameter space. If this is true, then for almost all admissible σ, τ and s, we have $\mathrm{Re}\, g_{21}(r_H) \neq 0$ and hence by Theorem 2 either supercritical or subcritical Hopf bifurcations exist.

In Figs. 1–5 we have plotted $\mathrm{Re}\, g_{21}(r_H)$ as a function of s, for several values of σ and τ. The values of σ and τ in Figs. 1–4 correspond to those in [5, Fig. 2], and the values of s at which $\mathrm{Re}\, g_{21}(r_H) = 0$ for rolls agree with [5]. In [2], Da Costa et al. give an analytic expression for $\mathrm{Re}\, g_{21}(r_H) = 0$, which also agrees with the results of [5]. The expression was obtained by considering a modal truncation of the full double-diffusive convection equations for rolls, which simplified the computation. However, the center manifold reduction used here provides justification for the validity of the truncation: enough modes were kept in [2] so as not to affect the normal form (3.2) in terms of order $\xi|\xi|^2$ or less.

We have also obtained results for squares and hexagons. In Figs. 1–4 the qualitative behavior of $\mathrm{Re}\, g_{21}(r_H)$ as a function of s for squares and hexagons is similar to the behavior for rolls: for s near s_C, $\mathrm{Re}\, g_{21}(r_H)$ is negative and increasing, and the graph crosses the s-axis as s is increased. The exact value of s at which $\mathrm{Re}\, g_{21}(r_H) = 0$ depends on the cellular structure considered. However, by varying σ and τ, one can observe

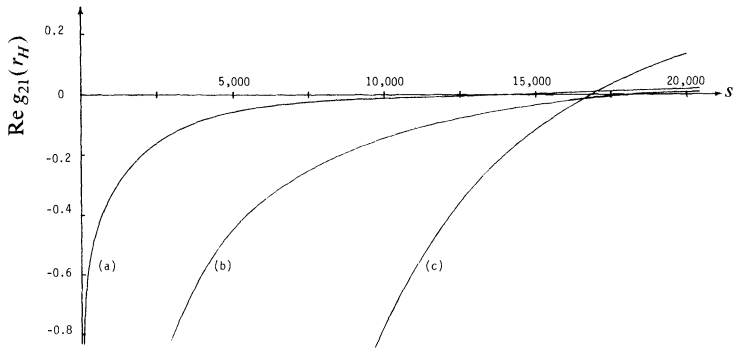


FIG. 1. Re $g_{21}(r_H)$ *as a function of* s *when* $\sigma = 1$, $\tau = 10^{-1/2}$ *for* (a) *rolls*, (b) *squares*, (c) *hexagons.*

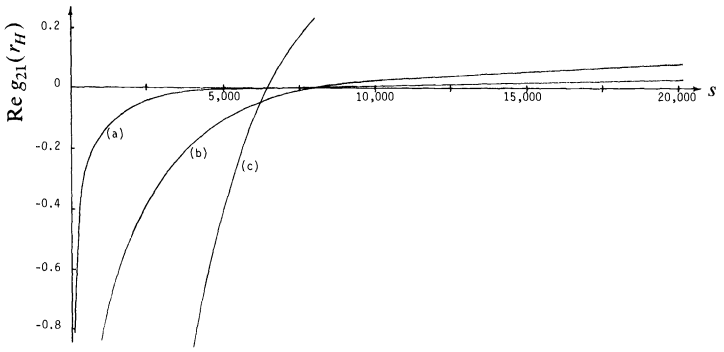


FIG. 2. Re $g_{21}(r_H)$ *as a function of* s *when* $\sigma = 1$, $\tau = 0.1$ *for* (a) *rolls*, (b) *squares*, (c) *hexagons.*

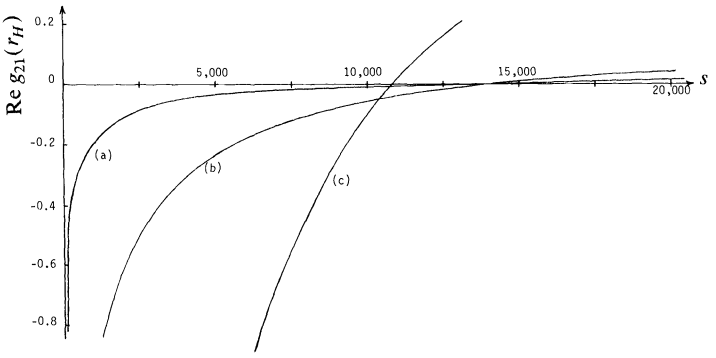


FIG. 3. Re $g_{21}(r_H)$ *as a function of* s *when* $\sigma = 10$, $\tau = 0.1$ *for* (a) *rolls*, (b) *squares*, (c) *hexagons.*

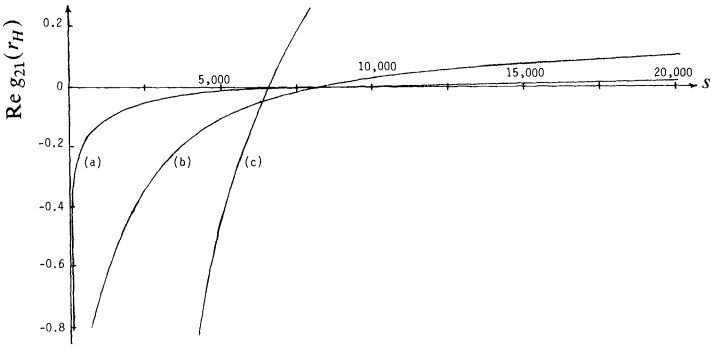


FIG. 4. Re $g_{21}(r_H)$ *as a function of* s *when* $\sigma = 7$, $\tau = 1/80$ *for* (a) *rolls*, (b) *squares*, (c) *hexagons.*

different behavior. For example, if the Prandtl number σ is decreased as in Fig. 5, where $\sigma = 0.001$ and $\tau = 0.1$, one sees that for s near s_C, $\mathrm{Re}\, g_{21}(r_H)$ is positive for squares and hexagons, but negative for rolls. Thus, branches of unstable time periodic square and hexagonal convection patterns bifurcate subcritically from the constant gradient solution at $r = r_H$, while branches of time periodic rolls bifurcate supercritically.

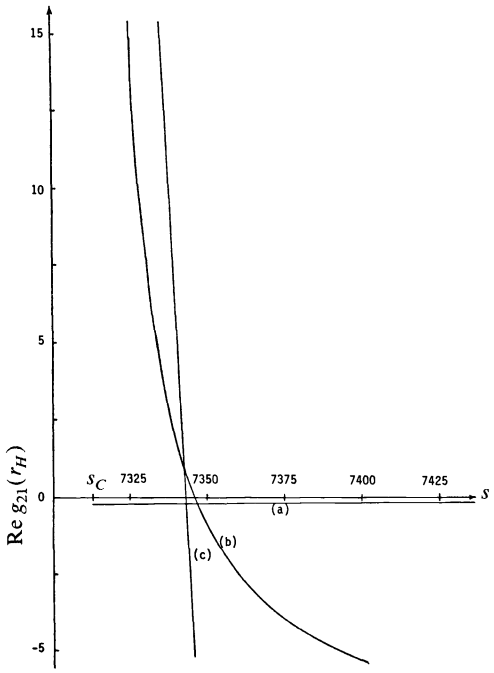


FIG. 5. $\mathrm{Re}\, g_{21}(r_H)$ *as a function of* s *when* $\sigma = 0.001$, $\tau = 0.1$ *for* (a) *rolls*, (b) *squares*, (c) *hexagons*.

It is possible that when $\mathrm{Re}\, g_{21}(r_H)$ is positive, the corresponding subcritical branch of unstable solutions may "turn back" at a limit point and the solutions on this branch beyond the limit point may be stable. This was conjectured by [5] for rolls, and the behavior of the modally truncated equations in [2] supports this conjecture. To verify this conjecture near parameter values where $\mathrm{Re}\, g_{21}(r_H) = 0$, one needs to compute the order $\xi|\xi|^4$ term in the normal form (3.2).

4. Conclusion. We have proved the existence of nontrivial time periodic solutions of double-diffusive convection equations, for suitable values of the parameters r, s, σ and τ. The solutions correspond to time periodic fluid flow in roll-like, rectangular, square and hexagonal convection cell patterns in the layer of fluid. The time periodic solutions bifurcate from the constant gradient solution either supercritically or subcritically, according to whether the sign of the coefficient $\mathrm{Re}\, g_{21}(r_H)$ is negative or positive. Solutions on a supercritical branch for r near r_H are orbitally asymptotically stable with respect to perturbations of the same cellular structure, while the solutions on a subcritical branch for r near r_H are unstable. As far as we know, a theory of pattern selection for Hopf bifurcation (which would determine whether solutions on a supercritical branch are also stable with respect to perturbations of a different cellular structure) is not yet available.

We have computed $\mathrm{Re}\, g_{21}(r_H)$ for rolls, squares and hexagons. For some parameter values, all three cellular structures exhibit similar functional dependence of $\mathrm{Re}\, g_{21}(r_H)$ on s, but as the Prandtl number σ is decreased, the behavior of $\mathrm{Re}\, g_{21}(r_H)$ for squares and hexagons is different than that for rolls.

REFERENCES

[1] M. G. CRANDALL AND P. H. RABINOWITZ, *The Hopf bifurcation theorem in infinite dimensions*, Arch. Rat. Mech. Anal., 67 (1977), pp. 53–72.

[2] L. N. DA COSTA, E. KNOBLOCH AND N. O. WEISS, *Oscillations in double-diffusive convection*, J. Fluid Mech., 109 (1981), pp. 25–43.

[3] B. D. HASSARD, N. D. KAZARINOFF AND Y.-H. WAN, *Theory and Applications of Hopf Bifurcation*, Cambridge Univ. Press, Cambridge, 1981.

[4] D. HENRY, *Geometric Theory of Semilinear Parabolic Equations*, Lecture Notes in Mathematics, 840, Springer-Verlag, Berlin, 1981.

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

[6] G. IOOSS, *Existence et stabilité de la solution périodique secondaire intervenant dans les problèmes d'evolution du type Navier-Stokes*, Arch. Rat. Mech. Anal., 47 (1972), pp. 301–329.

[7] V. I. IUDOVICH, *The onset of auto-oscillations in a fluid*, Prikl. Mat. Meh., 35 (1971), pp. 638–655; J. Appl. Math. Mech., 35 (1971), pp. 587–603.

[8] D. D. JOSEPH AND D. H. SATTINGER, *Bifurcating time periodic solutions and their stability*, Arch. Rat. Mech. Anal., 45 (1972), pp. 79–109.

[9] J. E. MARSDEN AND M. MCCRACKEN, *The Hopf Bifurcation and Its Applications*, Applied Mathematical Sciences Vol. 19, Springer-Verlag, New York, 1976.

[10] W. NAGATA AND J. W. THOMAS, *Bifurcation in double-diffusive systems* I. *Equilibrium solutions*, this Journal, this issue, pp. 91–113.

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

SIAM J. MATH. ANAL.
Vol. 17, No. 1, January 1986

013

THE GENERALIZED INVERSE OF AN UNBOUNDED LINEAR OPERATOR*

ROBERT NEFF BRYAN†

Abstract. Some necessary and some sufficient conditions are given for the existence of the Moore–Penrose generalized inverse of an unbounded linear operator between inner product spaces with an appropriately restricted domain. These results generalize a recent result of W. F. Langford.

1. Introduction. In this paper, we establish results similar to, but more general than, some of those given in a recent paper [1] by W. F. Langford regarding the existence of the Moore–Penrose generalized inverse of a certain unbounded linear operator between inner product spaces. Langford applied his results to a two-point ordinary differential boundary value problem to show that the associated operator has a generalized inverse, and, apparently, chose his assumptions on his operator so that his results would be applicable to the differential operator. Here, we show that his assumptions may be relaxed to some extent to obtain similar results, and we clarify the roles which certain of his assumptions play in the establishing of the results.

2. Notation and assumptions. We begin by listing our notation and stating our assumptions. Let X and Y denote real inner product spaces and l be a linear operator from X and Y. Let D be a subspace of X. We study the operator L which is l restricted to D. We denote the range of L by R, the kernel of l by k, and the kernel of L by K. The spaces of linear (not necessarily continuous) functionals on X and Y are denoted by X^* and Y^* respectively. The natural imbedding of X into X^* (and Y into Y^*) is denoted by J and maps $z \in X$ to $z^* \in X$ which satisfies $z^*(x) = \langle x, z \rangle$ for $x \in X$, where $\langle \cdot, \cdot \rangle$ is the inner product notation in X (and Y). We make extensive use of the following sets related to $S \subseteq X$:

$$S^{\perp} = \{ x \in X : \langle s, x \rangle = 0 \text{ for } s \in S \}$$

and

$$S^0 = \{ x^* \in X^* : x^*(s) = 0 \text{ for } s \in S \}.$$

The following lemma is interesting as it stands as well as useful in the sequel.

LEMMA 1. *Let X be an inner product space and S be a subspace of X. Then*

(1) *$J(S^{\perp}) = S^0$ implies $S^{\perp\perp} = S$;*

(2) *if $S^{\perp}$ is finite-dimensional and $S^{\perp\perp} = S$, then $J(S^{\perp}) = S^0$; and*

(3) *if $S^0 \subseteq J(X)$, then $J(S^{\perp}) = S^0$.*

Proof. (1) We need only show that $S^{\perp\perp} \subseteq S$. For some $t \in S^{\perp\perp} - S$ there is a functional $u \in X^*$ for which $u(t) = 1$ and $u(s) = 0$ for $s \in S$; hence, there exists $\bar{u} \in S^{\perp}$ for which $J(\bar{u}) = u$. This implies $\langle \bar{u}, t \rangle = 1$ which contradicts that $t \in S^{\perp\perp}$ and $u \in S^{\perp}$.

(2) Clearly, $J(S^{\perp}) \subseteq S^0$. We have, by hypothesis, $X = S \oplus S^{\perp}$. Let $u \in S^0$, and consider the restriction of u to $S^{\perp}$. Since $S^{\perp}$ is finite-dimensional, there is a $\bar{u} \in S^{\perp}$

*Received by the editors August 11, 1981, and in revised form March 19, 1984.

† Department of Mathematics, University of Western Ontario, London, Ontario, Canada N6A 589.

satisfying $u(x)=\langle \bar{u}, x\rangle$ for $x\in S^{\perp}$. Given an $x=x_1+x_2$, $x_1\in S$, $x_2\in S^{\perp}$, $u(x)=u(x_2)$ $=\langle \bar{u}, x_2\rangle=\langle \bar{u}, x\rangle$. Hence, $u=J(\bar{u})$ and $u\in J(S^{\perp})$. Thus, $J(S^{\perp})=S^0$.

(3) The proof is immediate.

3. Main results. First, we review some terminology. The equation $Lx=y$ has associated with it the set $S_y=\{x\in X:\ \|Lx-y\|\leqq\|Lz-y\|;\ z\in X\}$ of least squares solutions. The norm is given by the inner product on Y. The set S_y is convex. If there is a unique point x^+ in S_y satisfying $\|x^+\|\leqq\|x\|$ for all $x\in S_y$, then x^+ is called the best-least-squares solution (BLSS) of $Lx=y$. The operator L is said to have a (Moore–Penrose) generalized inverse, L^+, if $Lx=y$ has a BLSS for every y in Y, and, of course, $L^+y=x^+$.

In order to compare the results which follow with those of Langford in [1], the reader should keep in mind that in the setting of [1], the spaces k (and hence K) and $R^{\perp}$ are finite-dimensional. The assumptions regarding each of these spaces in this paper are included in each statement of a lemma or theorem which involves it, except Theorem 4 in which $R^{\perp}$ is assumed to be finite-dimensional.

Advantages of considering these slightly more general settings are that some proofs are simplified and the roles that the various assumptions play in the results are made explicit. This enables us to concentrate on the sets of sufficient conditions for the existence of L^+, each of which includes the completeness of K; they appear as the first statement in each of Theorems 1, 2, and 3 and in Theorem 4.

The finite-dimensional version of our first main result does not appear in [1], but is hinted at in the proof of Langford's Lemma 1 [1, p. 1085].

THEOREM 1. (1) *If $R^{\perp}$ and K are complete and $R^{\perp\perp}=R$, then L^+ exists*; (2) *if $R^{\perp}$ is complete and L^+ exists, then $R^{\perp\perp}=R$.*

Proof. (1) Since $R^{\perp}$ is complete, $Y=R^{\perp\perp}\oplus R^{\perp}=R\oplus R^{\perp}$. Hence, y, in Y, can be written as $y=y_1+y_2$, $y_1\in R$, $y_2\in R^{\perp}$. The projection theorem implies that $\|y-y_1\|<$ $\|y-z\|$ for all $z\in R-y_1$. Letting x_1 be a solution of $Lx=y_1$, we see that the inequality implies $\|Lx_1-y\|\leqq\|Lx-y\|$ for all $x\in D$. Since K is a complete subspace of X, we can apply the projection theorem again to obtain unique $m_0\in K$ satisfying $\|x_1-m_0\|\leqq\|x_1-m\|$ for all $m\in K$. Then $x^+=x_1-m_0$ is the BLSS of $Lx=y$. Thus, L^+ exists.

(2) Let $y\in Y$. There exists a unique $x^+=L^+u$ which satisfies $\|Lx^+-y\|\leqq\|Lx-y\|$ for all $x\in D$, so that there is a unique $\bar{y}=Lx^+\in R$ such that $\|\bar{y}-y\|\leqq\|z-y\|$ for all $z\in R$. Thus, by the alternate form of the projection theorem (see §6), $Y=R\oplus R^{\perp}$. But $Y=R^{\perp}\oplus R^{\perp\perp}$, so that $R^{\perp\perp}=R$.

The following theorem is a generalization of Langford's lemma [1, p. 1085], mentioned above. We note, in particular, that no assumptions are made here on the space k. The proof consists of applying Lemma 1 and Theorem 1.

THEOREM 2. (1) *If $R^{\perp}$ and K are complete and $J(R^{\perp})=R^0$, then L^+ exists.* (2) *If $R^{\perp}$ is finite-dimensional and L^+ exists, then $J(R^{\perp})=R^0$.*

The adjoint of l, denoted by l^*, is defined in the usual manner. If k, the kernel of l, is complete, then $R(l^*)$, the range of l^*, is the annihilator of k; i.e., $R(l^*)=k^0$. A proof of this is given in [1, Lemma 8, p. 1094], where only the completeness of k is used, not its finite-dimensionality.

We note that the following result does not depend on the dimensionality of any of the spaces involved.

LEMMA 2. (cf. [1, Lemma 2, p. 1086]). (1) $l^*(R^0)\subseteq D^0\cap k^0$; (2) *furthermore, if k is complete, then $D^0\cap k^0=l^*(R^0)$.*

Proof. (1) Let $u \in l^*(R^0)$, $v \in R^0$ so that $u = l^*v$. Then $u(x) = l^*v(x) = v(lx) = 0$ for $x \in D$, so $u \in D^0$. Clearly $u(x) = 0$ for $x \in k$, so $u \in k^0$.

(2) For $u \in D^0 \cap k^0$, $u \in R(l^*)$ so there is a $v \in Y^*$ for which $u = l^*$. Since $u \in D^0$, $l^*v(x) = 0$ for every $x \in D$. But $l^*(v) = v(lx)$ so $v(r) = 0$ for every $r \in R$; i.e., $v \in R^0$, so $u \in l^*(R^0)$. Thus, $D^0 \cap k^0 \subset l^*(R^0)$. Part (1) yields the equality,

Theorems 3 and 4, below, are generalizations of [1, Thm. 1]. The statements of those theorems indicate the interplay among the various assumptions made on k, K, and $R^\perp$ which are lost in [1, Thm. 1]. Together with Theorems 1 and 2, above, they provide several sets of necessary conditions and of sufficient conditions for the existence of L^+.

THEOREM 3. (1) *If $R^\perp$, k, and K are complete and $D^0 \cap k^0 \subseteq l^*[J(Y)]$, then L^+ exists.*

(2) *If L^+ exists, $R^\perp$ is finite-dimensional, and k is complete, then $D^0 \cap k^0 \subseteq l^*[J(Y)]$.*

Proof. (1) We have $D^0 \cap k^0 = l^*(R^0) \subseteq l^*[J(Y)]$ by Lemma 2, so that $R^0 \subseteq J(Y)$, since l^* is one-to-one. Thus, $R^0 = J(R^\perp)$, by Lemma 1, so L^+ exists, by Theorem 2.

(2) We use Lemma 2, Theorem 2, and the fact that $R^\perp \subseteq Y$, in turn, to establish that $D^0 \cap k^0 = l^*(R^0) = l^*[J(R^\perp)] \subset l^*[J(Y)]$.

The set $l^*[J(Y)]$, which is mentioned in Theorem 3, has an attractive characterization, which adds interest to that theorem. The characterization is given in the following remark.

Remark. If $u \in l^*[J(Y)]$, then there is a point $y \in Y$ for which $u(x) = \langle lx, y \rangle$ for $x \in X$, and conversely.

Using this result, we see, readily, the following lemma.

LEMMA 3. *If $D^0 \subseteq J(k) \oplus l^*[J(Y)]$, then $D^0 \cap k^0 \subseteq l^*[J(Y)]$.*

This gives us a sufficient condition on D for the existence of L^+; viz.,

COROLLARY (cf. [1, Corollary 1, p. 1086]). *If $R^\perp$, k, and K are complete and $D^0 \subseteq J(k) \oplus l^*[J(Y)]$, then L^+ exists.*

4. A result for L with special domain. In this section we choose a particular type of domain of L, essentially the same as that described in [1], and obtain a necessary and sufficient condition for the existence of L^+.

Let $\{f_1, f_2, \cdots, f_m\}$ be a linearly independent set in X^* and F be the subspace of X^* spanned by $f_1, \cdots, f_m$. Let $D = \{x \in X: f(x) = 0 \text{ for all } f \in F\}$. It is easily seen that D, given in this way, has finite codimension and that $D^0 = F$. This implies that $R^\perp$ is finite-dimensional and, hence, complete. Thus, for D as given here, we have, from Theorem 3, the following theorem.

THEOREM 4. *If k and K are complete, then L^* exists if and only if $F \cap k^0 \subseteq l^*[J(Y)]$.*

5. An example. We include an example to show that Theorem 1 applies in at least one case where Langford's Theorem 1 does not. Let $X_1 = \{a \in l^2: a_n = 0 \text{ for sufficiently large } n\}$ with l^2 inner product, $X = l^2 \oplus^\perp X_1$ with inner product $\langle s_1 + x_1, s_2 + x_2 \rangle = \langle s_1, s_2 \rangle + \langle x_1, x_2 \rangle$, and $Y = X_1$. Let T: $X_1 \to Y$ be defined by $T(a) = T[(a_1, a_2, \cdots)] = (a_1, 2a_2, 3a_3, \cdots)$. Then T is one-to-one but not continuous. Define l by $l(x) = l(s + a) = T(a)$, and let $D = \theta \oplus X_1$. Note that D has infinite codimension. Let L be l restricted to D. Then $R = Y = X_1$, $R^\perp = \{\theta\}$, which is complete, and $R^{\perp\perp} = R$; also, L is not continuous. The kernels, $k = K = l^2 \oplus \theta$, are infinite-dimensional but complete. Thus, L has a generalized inverse, by Theorem 1.

6. The projection theorem. We include two versions of the projection theorem which were applied in the proof of Theorem 1.

PROJECTION THEOREM. *Let V be an inner product space, M be a subspace of V and $v \in V - M$ and $m_0 \in M$. Then $\|v - m_0\| < \|v - m\|$ for all $m \in M - m_0$ if and only if $v - m_0 \in M^{\perp}$.*

PROJECTION THEOREM–ALTERNATE FORM. *Let V be an inner product space and M a subspace of V. Then for every $v \in V$ there exists a unique $m_0 \in M$ such that $\|v - m_0\| \leqq \|v - m\|$ for all $m \in M$ if and only if $V = M \oplus M^{\perp}$.*

REFERENCES

[1] W. F. LANGFORD, *The generalized inverse of an unbounded linear operator with unbounded constraints*, this Journal, 9 (1978), pp. 1083–1095.

SIAM J. MATH. ANAL.
Vol. 17, No. 1, January 1986

014

LINEARIZATION STABILITY FOR AN INVERSE PROBLEM IN SEVERAL-DIMENSIONAL WAVE PROPAGATION*

W. W. SYMES†

Abstract. Inverse problems in wave propagation arise from the physical notion that the response of a mechanical continuum to a specified excitation should reflect the properties of the medium through which the excited waves travel, even though the waves are eventually measured at a remote location. We consider the formal linearization of a simple model inverse problem in several-dimensional wave propagation, in which the density distribution of a linear fluid is to be recovered from its remotely measured response to an incident plane-wave excitation, assuming the sound velocity distribution to be known and constant. We make reasonable choices for norms (error measures), and give simple examples establishing the ill-posed nature of the problem and indicating the mechanism of instability. We then show how the problem may be regularized (rendered well-posed) by the introduction of minimally stringent a priori constraints, motivated by a (crude) approximation to the singular value decomposition of the linearized problem.

AMS-MOS subject classification (1980). Primary 35R25

Key words. inverse, problem in wave propagation, linearization stability, regularization

Inverse problems in wave propagation arise from the physical notion that the response of a mechanical continuum to a specified excitation should reflect the properties of the medium through which the excited waves travel, even though the waves are eventually measured at a remote location. All measurement is contaminated by error, and the mathematical structures used to model wave propagation can be regarded only as approximate descriptions of actual physical processes. Therefore, the most important question about these problems (as for many other problems in applied mathematics) is: to what extent does error in data (remote measurement of wave fields) produce error in solution (estimates of mechanical parameters)? This question of *stability* or *condition* is also critical for the design of effective numerical algorithms.

In this paper, we address a very simple model inverse problem in several-dimensional wave propagation, by examining its formal linearization. We make (reasonable) explicit choices for norms (error measures) and give simple examples establishing the ill-posed nature of the linearized problem, and indicating the source of the difficulty. We then show how the problem may be rendered well-conditioned (*regularized*) by imposition of a priori constraints on the mechanical parameters. These constraints are essentially the least stringent possible, and amount to a crude approximation to the singular value decomposition of the linearized problem.

We expect these considerations to apply with some changes to more complex inverse problems of greater practical interest than the simple model problem considered here.

We consider an acoustic medium confined to a half-space $\{x \in \mathbb{R}^n: x_n \geqq 0\}$. We assume that the excess pressure u is small, and so is governed approximately by the linear acoustic wave equation

$$\left(\frac{1}{\rho c^2}\partial_t^2 - \nabla \cdot \frac{1}{\rho}\nabla\right) u = 0 \tag{1}$$

* Received by the editors March 15, 1984, and in revised form March 1, 1985. This research was supported in part by the National Science Foundation under grant MCS-80-02996-01, and by the Office of Naval Research under contract N00014-83-K-0051.

† Department of Mathematical Sciences, Rice University, Houston, Texas 77001.

where ρ is the density and c the sound velocity; both depend on x. We assume that the fluid is initially quiescent, i.e., $u \equiv 0$ for $t \ll 0$, and that the boundary of the fluid is *fixed* (rigid) except at $t = 0$, when it is subject to an impulsive acceleration, uniformly over the entire boundary:

$$\partial_{x_n} u(\cdot, 0, t) = -\delta(t) \tag{2}$$

where we have written $x = (x', x_n)$, $x' \in \mathbb{R}^{n-1}$. A final simplifying assumption is that the velocity $c \equiv 1$, so that the rays of geometric acoustics are straight lines.

We take for the data of our problem the excess pressure measured at the boundary $x_n = 0$. Since the problem (1) is well posed, at least for smooth ρ, each distribution of density produces a (hopefully characteristic) response

$$\mathcal{T}(\rho) = u(\cdot, 0, \cdot).$$

The inverse problem thus amounts to the study of the functional equation

$$\mathcal{T}(\rho) = g. \tag{3}$$

Instead of considering (3) directly, we will examine its formal linearization

$$D\mathcal{T}(\rho^0) \cdot \rho^1 = g^1 \tag{4}$$

in which a perturbation ρ^1 about a reference density distribution ρ^0 is sought to produce a perturbation g^1 in the response.

We first give an example which shows that, for reasonable choices of norms (error measures) for ρ and g, one or both of $D\mathcal{T}$ and $D\mathcal{T}^{-1}$ are unbounded. This means that:

(i) Depending on choice of norm, either $\mathcal{T}$ is not differentiable, or its derivative is not boundedly invertible, or $D\mathcal{T}$ is not its derivative, or any sensible combination of these circumstances.

(ii) The equation (4) does not in general possess solutions.

Consequently, the only obvious computational option for (3), i.e. some version of Newton's method, is doomed to failure. Problem (3) and (4) must be modified by the introduction of a priori constraints, to restore continuous dependence of ρ^1 on g^1. By virtue of (ii), we cannot actually expect to solve (4), so we change it to a least-squares optimization problem with constraint, i.e. we seek to minimize

$$\left\| D\mathcal{T}(\rho^0) \cdot \rho^1 - g^1 \right\|^2, \qquad \rho^1 \in \mathfrak{R} \tag{5}$$

where $\mathfrak{R}$ is a *regularizing set* and the norm is an appropriate Hilbert space norm. If the optimization problem (5) is to have stable solutions, an inequality of the form

$$\|\rho^1\| \leqq C \left\| D\mathcal{T}(\rho^0) \cdot \rho^1 \right\|, \qquad \rho^1 \in \mathfrak{R} \tag{6}$$

must hold. The derivation of roughly such an inequality is the main goal of this paper.

The choice of $\mathfrak{R}$ is important. Unfortunate choices of $\mathfrak{R}$ may override some information present in the data. We want to choose $\mathfrak{R}$ to correspond roughly to the *small singular values* of $D\mathcal{T}$: that is, we want the condition $\rho^1 \in \mathfrak{R}$ to bound a priori the components of ρ^1 in the singular value decomposition corresponding to small singular values, but leave the other components unbounded. We identify a suitable $\mathfrak{R}$ below (though our choice is probably suboptimal, as we also discuss).

To complete the picture, and deduce similar results for the nonlinear problem (3), we need to establish that $\mathscr{T}$ is actually differentiable, and that $D\mathscr{T}$ is continuous. These goals can be accomplished by extending the arguments given in our previous paper [12], in which we established that $\mathscr{T}$ is Lipshitz-continuous when the domain (ρ) is metrized with an anisotropic Sobolev norm. The results of [12] are fundamental to those presented here, and we make the same choice of norm for the perturbations ρ^1 below.

To end this introduction, we note that essentially the same arguments yield similar results for *velocity-inversion*, a more practically motivated problem in which the sound velocity is to be determined from remote pressure measurements, provided that no caustics occur in the incident wavefront. This restriction is highly unnatural, but the presence of caustics introduces presently unresolved technical difficulties. On the other hand, many authors have discussed (formal aspects of) linearized inverse problems, most often linearized velocity inversion against a homogeneous or stratified background—see for instance [3] and references cited therein. Also, most current data-processing methods in seismic exploration for petroleum, nondestructive materials evaluation by ultrasonic probing, ocean acoustics, etc., are based in one way or another on the study of formal linearizations. Some authors have also written on iterative, formal algorithms for the solution of inverse problems of this type—see [8], [9], [10], [11]. For one-dimensional problems, a fair amount about the analytic structure of analogues of $\mathscr{T}$ is known—see [13], [15] and references cited there. In higher dimensions, very little of a nonformal nature seems to have been accomplished.

We introduce some convenient changes in notation: we write: $x \to (x,z)$, so that $x \in \mathbb{R}^{n-1}$, $z \in \mathbb{R}^+$. We use the notation $\|u\|_s$ for the norm of $u \in H^s(\Omega)$, suppressing the dependence on the domain Ω. We set

$$\partial_t = \frac{\partial}{\partial t}, \quad \partial_z = \frac{\partial}{\partial z}, \qquad D_k = \frac{\partial}{\partial x_k}, \quad k = 1, \cdots, n-1,$$

$$\nabla_x = (D_1, \cdots, D_{n-1}), \qquad \nabla = (\nabla_x, \partial_z),$$

$$\nabla^2 = \Delta = \sum_{j=1}^{n-1} \frac{\partial^2}{\partial x_j^2} + \frac{\partial^2}{\partial z^2} = \Delta_x + \frac{\partial^2}{\partial z^2}.$$

We shall also replace ρ by η^{-1} for convenience. Denote by u the solution of (1), i.e.

$$(1') \qquad \begin{aligned} &\left(\partial_t^2 - \Delta - \nabla \log \eta \cdot \nabla\right) u = 0, \\ &\partial_z u(\cdot, 0, t) = -\delta(t) \\ &u \equiv 0, \qquad t \ll 0. \end{aligned}$$

We suppose throughout that $\log \eta \in C^\infty(\mathbb{R}^n_+)$. It is then standard that u is smooth in the forward light cone $C_\infty = \{(x,z,t)\colon t > z > 0\}$, and on the boundary (wave front) $\{z = t\}$ undergoes a jump discontinuity determined by the transport equation of geometric optics. Consequently, inside the light cone u solves the characteristic initial/boundary value problem

$$(7) \qquad \begin{aligned} &\left(\partial_t^2 - \Delta - \nabla \log \eta \cdot \nabla\right) u \equiv 0, \\ &\partial_z u(\cdot, 0, \cdot) \equiv 0, \qquad t > 0, \\ &u(x, z, z^+) = \eta(x,0)^{1/2} \eta(x,z)^{-1/2} \end{aligned}$$

(see [5, pp. 42–46] or [4, Chap. VI, §4, pp. 633–655]).

Remark. The transport equation breaks down in the presence of caustics, which complicates the extension of the present reasoning to variable-velocity problems.

The characteristic geometry of the wave equation shows that $\{\eta(x,z)\colon z\leqq T\}$ determines $\{u(x,0,t)\colon t\leqq 2T\}$. For $\log\eta\in C^\infty(\mathbb{R}^{n-1}\times[0,T])$ or $C^\infty(\mathbb{R}^n_+)$ respectively, we define

$$\mathscr{T}_T(\eta) := u(\cdot,0,\cdot)\in C^\infty\big(\mathbb{R}^{n-1}\times[0,2T]\big),$$
$$\mathscr{T}_\infty(\eta) := u(\cdot,0,\cdot)\in C^\infty\big(\mathbb{R}^n_+\big).$$

When the value of T is implicit from context, we suppress the subscript, and write $\mathscr{T}_T=\mathscr{T}$.

We next formally linearize $\mathscr{T}$. Let $\eta=\eta^0+\varepsilon\eta^1$, $u=u^0+\varepsilon u^1+\cdots$, and substitute in (7). A short calculation yields the perturbational equations

$$\begin{aligned}&\big(\partial_t^2-\Delta-\nabla\log\eta^0\cdot\nabla\big)u^1=\nabla\frac{\eta^1}{\eta^0}\cdot\nabla u^0,\\ &\partial_z u^1(\cdot,0,\cdot)\equiv 0,\\ &\bar u^1=-\frac{1}{2}\eta_0^{1/2}\big(\eta^0\big)^{-3/2}\eta^1\end{aligned}\tag{8}$$

where we have used the useful notation $\bar u(x,z)=u(x,z,z^+)$. Thus

$$D\mathscr{T}\big(\eta^0\big)\cdot\eta^1=u^1(\cdot,0,\cdot).$$

To discuss the behavior of $D\mathscr{T}$, we must introduce norms for η^1 and g^1. For reasons discussed below, we choose the H^1-norm for η^1, and the Dirichlet norm for g^1.

We now sketch a pair of examples, based on the geometric optics construction, which show that $D\mathscr{T}(1)$ and $D\mathscr{T}(1)^{-1}$ are both unbounded. A very similar argument is given in [14, §3] for the time-like Cauchy problem, so we refer the reader to that paper for details of the estimates.

In fact these examples really concern the relation between initial and boundary values for solution of the wave equation, and express well-known facts. The first example (unboundedness of $D\mathscr{T}^{-1}$) shows (essentially) the ill-posedness of the time-like Cauchy problem. The second (unboundedness of $D\mathscr{T}$) shows that the boundary values are generally not as smooth as initial data. Perhaps by sharpening the second example, one could show that, in general, one loses exactly the half-derivative suggested by the trace theorem.

The first example also follows trivially from the existence of solutions with singularities along a single ray—see [17]. Presumably the second example could also be derived from a closer examination of the Gaussian beam construction, although the route pursued below (ordinary geometric optics) seems simpler. In both cases the pathology arises from energy propagating along grazing rays.

The simplest reference coefficient to consider is $\eta^0=1$, for which

$$u^0(x,z,t)=h(t-z)$$

where h is the Heaviside unit step function.

For u^1, obtain the equations for the so-called Born approximation:

$$\begin{aligned}&\big(\partial_t^2-\nabla^2\big)u^1=0, \qquad t>z,\\ &\partial_z u^1(x,0,t)=0, \qquad t>0,\\ &u^1(x,z,z^+)=-\tfrac{1}{2}\eta^1(x,z).\end{aligned}\tag{9}$$

For convenience, we temporarily allow η^1 and u^1 to be complex, with the understanding that only the real parts have significance.

Suppose χ is a smooth function of bounded support in $\mathbb{R}^2$ with unit L^2 norm, and set

$$\eta^1(x,z)=\frac{1}{i\omega}\chi(x,z)e^{i\omega(x+z)},$$

$$u_a^1(x,z,t)=\frac{-1}{2i\omega}\chi(x-z+t,z)e^{i\omega(x+t)} \quad \text{for } t>z.$$

In fact, u_a^1 is a horizontally moving monochromatic wave packet, and is an approximate solution of (9) in the sense that

$$(\partial_t^2-\nabla^2)u_a^1=O\left(\frac{1}{\omega}\right).$$

Also

$$u_a^1(x,z,z^+)=-\tfrac{1}{2}\eta^1(x,z).$$

If we reflect u_a^1 at the boundary i.e., replace $u_a^1(x,z,t)$ by $u_a^1(x,z,t)+u_a^1(x,-z,t)$, then u_a^1 differs from u^1, in the energy norm, by $O(1/\omega)$.

We distinguish two cases. First suppose that the support of χ is entirely contained in $\{z>0\}$. Then $u_a^1\equiv 0$, for $z=0$; it follows that there is some sequence of frequencies $\omega_n\to\infty$ for which

$$\int_0^T\int_{\mathbb{R}^{n-1}}dx\left\{\left|\partial_t u^1(x,0,t)\right|^2+\left|\nabla_x u^1(x,0,t)\right|^2\right\}\to 0$$

while

$$\int dz\int\left|\nabla\eta^1\right|^2\to 1.$$

That is, the boundary values of u^1 can become arbitrarily small in the mean-square sense, even when the derivatives of the coefficient η^1 remain large. Physically, the level surfaces of the oscillatory perturbation η^1 dip at 45°, and reflect the incident vertical rays horizontally. Since for large ω the field is essentially given by its geometric optics approximation u_a^1, the trace becomes small as $\omega\to\infty$.

For our second example, we assume that the support of χ does intersect $\{z=0\}$. We construct a family of η_ε^1, $u_{a,\varepsilon}^1$ depending on another parameter $\varepsilon>0$ as follows:

$$\eta_\varepsilon^1(x,z)=\frac{1}{i\omega}\chi\left(x,\frac{z}{\varepsilon}\right)e^{i\omega(x+z)},$$

$$u_{a,\varepsilon}^1(x,z,t)=\frac{-1}{2i\omega}\chi\left(x-z+t,\frac{z}{\varepsilon}\right)e^{i\omega(x+t)}.$$

Then

$$(\partial_t^2-\nabla^2)u_{a,\varepsilon}^1=O\left(\frac{1}{\omega\varepsilon^2}\right).$$

Also

$$\iint\left|\nabla\eta_\varepsilon^1\right|^2dx\,dz=O(\varepsilon)+O\left(\frac{1}{\omega^2\varepsilon}\right).$$

Now let $\omega \to \infty$, $\varepsilon \to 0$ in such a way that $\omega\varepsilon^2 \to \infty$. Then, as in [14], we can show that for some sequence of values of ω, ε,

$$\int \int_0^T \left|\partial_t u^1_{a,\varepsilon} - \partial_t u^1\right|^2_{(z=0)} dx\, dt \to 0.$$

However,

$$\int dx \int_0^T dt \left|\partial_t u^1_{a,\varepsilon}\right|^2_{(z=0)} = O(1)$$

whereas

$$\iint dx\, dz \left|\nabla \eta^1_\varepsilon\right|^2 \to 0.$$

This second example shows that $D\mathscr{T}$ itself is unbounded. As the inverse is also unbounded by the first part, no strengthening or weakening of topologies of the domain and range can render both bounded. We shall change the topology in the domain so as to make $D\mathscr{T}$ bounded, then ask for optimal regularization of $D\mathscr{T}^{-1}$.

Note that in both cases, the cause of unboundedness was rapid oscillation of η^1 in the x-("horizontal") directions. Therefore it seems reasonable that imposition of additional smoothness in horizontal directions might cure the difficulty, and this is indeed the case.

To state our main results, we shall make use of the spaces $\mathscr{H}_{(m,s)}$ [7, §2.5]. These are the completions of $C_0^\infty(\mathbb{R}^n)$ in the norms ($s \geqq 0$)

$$\|u\|^2_{(m,s)} = \sum_{j=0}^{m} \left\|\partial_z^j u\right\|^2_{(0,s+m-j)},$$

$$\|u\|_{(0,s)} = \left\|(1-\Delta_x)^{s/2} u\right\|_{L^2(\mathbb{R}^n)}.$$

The following facts are either easy to prove or are stated explicitly in [7, §2.5]:

(I) For any multiindex $\alpha = (\alpha', \alpha_z)$, $D^\alpha u \in \mathscr{H}_{(p,t)}$ if $u \in \mathscr{H}_{(m,s)}$ and $|\alpha| < m+s-(p+t)$, $\alpha_z \leqq m-p$.

(II) If $m+s > n/2$ and $m > \frac{1}{2}$, then $\mathscr{H}_{(m,s)} \subset L^\infty(\mathbb{R}^n) \cap C^0(\mathbb{R}^n)$ (continuous inclusion).

(III) $\|\ \|_{(1,s)}$ is equivalent to the graph norm of $(-\Delta_x)^{s/2}$, viewed as a densely defined operator on $H^1(\mathbb{R}^n)$.

THEOREM 1. *For $s > (n-1)/2$, there exists a constant $C > 0$ depending on s, T, n, $\|\log \eta^0(\cdot,0)\|_{H^{s+1}(\mathbb{R}^{n-1})}$, and $\|\log \eta^1\|_{\mathscr{H}_{(1,s+2)}(\mathbb{R}^{n-1}\times[0,T])}$ so that for η^0, $\eta^1 \in C^\infty(\mathbb{R}^n_+)$, $\|D\mathscr{T}_T(\eta^0)\cdot\eta^1\|_1 \leqq C\|\eta^1\|_{(1,s+2)}$. Also for η^0, $\tilde{\eta}^0$, $\eta^1 \in C^\infty(\mathbb{R}^n_+)$,*

$$\left\|D\mathscr{T}_T(\eta^0)\cdot\eta^1 - D\mathscr{T}_T(\tilde{\eta}^0)\cdot\eta^1\right\|_1 \leqq C\left\|\log\eta^0 - \log\tilde{\eta}^0\right\|_{(1,s+2)}\left\|\eta^1\right\|_{(1,s+2)}.$$

In particular, $D\mathscr{T}_T$ extends to a Lipschitz-continuous map

$$\log\eta^0 \to D\mathscr{T}_T(\eta^0),$$

$$\mathscr{H}_{(1,s+2)}(\mathbb{R}^{n-1}\times[0,T]) \to \mathscr{L}\left[\mathscr{H}_{(1,s+2)}(\mathbb{R}^{n-1}\times[0,T]), H^1(\mathbb{R}^{n-1}\times[0,2T])\right].$$

Suppose $\gamma > 0$, and suppose $\Gamma \subset \mathbb{R}^n_+$ is a closed set for which $\Gamma \cap \{z = z_0\}$ is compact for every $z_0 \geqq 0$. We call such a Γ laterally compact. Define

$$\mathfrak{R}(\gamma, \Gamma) = \Big\{ f \in H^1(\mathbb{R}^n_+) : \operatorname{supp} f \subset \Gamma$$
$$\text{and for every } T > 0,\ \|f\|_{\mathscr{H}_{(1,s+2)}(\mathbb{R}^{n-1} \times [0,T])} \leqq \gamma \|f\|_{H^1(\mathbb{R}^{n-1} \times [0,T])} \Big\}.$$

Remark. An inequality of the above form holds for every $f \in \mathscr{H}_{(1,s+2)}(\mathbb{R}^n_+)$ and every $T > 0$, but not necessarily with γ independent of T. Thus this inequality amounts to a strong form of the ban on horizontal oscillation mentioned above. If $f \in H^1(\mathbb{R}^n_+)$ and $\varphi \in C_0^\infty(\mathbb{R}^{n-1})$, then $f * \hat{\varphi}$ (convolution in the x-variables) obeys an inequality of this form. Also, results entirely analogous to those given here may be formulated for cylinder domains $\Omega \times \mathbb{R}_+$, rather than $\mathbb{R}^n_+ \simeq \mathbb{R}^{n-1} \times \mathbb{R}_+$, with Dirichlet or Neumann boundary conditions on $\partial\Omega \times \mathbb{R}_+$, or any other choice of boundary condition for which various integrations by points in the x-variables remain valid. In terms of the eigenfunctions of the $(n-1)$-dimensional Laplacian on such a (compact) Ω, an interesting class of subspaces of H^1 lying inside the regularizing sets $\mathfrak{R}$ are the H^1-functions of z with values in finite Fourier series in x, although these (linear) subspaces are clearly proper subsets of the conic regularizing sets $\mathfrak{R}$.

THEOREM 2. *Suppose T, $\gamma > 0$ and $\Gamma \subset \mathbb{R}^{n-1}$ is laterally compact. Then there exists a functional K: $\mathfrak{R}(\gamma, \Gamma) \to \mathbb{R}_+$ so that*

(i) *If $\{\eta^1_j\} \subset \mathfrak{R}(\gamma, \Gamma)$, $\eta^1_j \to 0$ weakly in $H^1(\mathbb{R}^n_+)$, then $K(\eta^1_j) \to 0$.*

(ii) *There exists $C > 0$ depending on $\|\log \eta^0\|_{(1,s+2)}$, s, n, and T, for which*

$$\|\eta^1\|^2_{H^1(\mathbb{R}^{n-1} \times [0,T])} \leqq C\Big(\big\| D\mathscr{T}_T(\eta^0) \cdot \eta^1 \big\|^2_{H^1(\mathbb{R}^{n-1} \times [0,2T])} + K(\eta^1) \Big).$$

Remark. The presence of the "compact" functional K on the r.h.s. of the above estimate means that we can only assert that a sufficiently small neighborhood of zero contains finitely many singular values of $D\mathscr{J}$. In particular we cannot rule out zero as a singular value (i.e. nonuniqueness), and therefore have not quite established continuity of the inverse. The best we can do in this direction is summarized in:

THEOREM 3. *Suppose that $\log \eta^0 \in \mathscr{H}_{(1,s+2)}(\mathbb{R}^n_+) \cap W^{1,\infty}(\mathbb{R}^n_+)$, and that $\gamma > 0$ and $\Gamma \subset \mathbb{R}^n_+$ is laterally compact. Then if $\eta^1 \in \mathfrak{R}(\gamma, \Gamma)$ satisfies*

$$D\mathscr{T}_\infty(\eta^0) \cdot \eta^1 = 0$$

it follows that $\eta^1 \equiv 0$.

THEOREM 4. *Suppose η^0, γ, Γ as above, and define*

$$\mathscr{A}: \mathfrak{R}(\gamma, \Gamma) \to H^1_{\mathrm{loc}}(\mathbb{R}^n_+) \oplus H^1(\mathbb{R}^{n-1})$$

by

$$\mathscr{A}\eta^1 = \big(D\mathscr{T}_\infty(\eta^0) \cdot \eta^1, \eta^1(\cdot, 0) \big).$$

Let $\mathscr{M}$ be a weakly closed subset of

$$\bigcup_{\gamma' > 0} \mathfrak{R}(\gamma', \Gamma)$$

and let $\mathscr{M}_\gamma = \mathscr{M} \cap \mathfrak{R}(\gamma, \Gamma)$, *given the topology of* $H^1_{\text{loc}}(\mathbb{R}^n_+)$. *Then*:

(i) *The restriction* $\mathscr{A}_\gamma$ *of* $\mathscr{A}$ *to* $\mathscr{M}_\gamma$ *has closed range.*

(ii) $\mathscr{A}_\gamma^{-1}$ *is continuous on* Range $\mathscr{A}_\gamma$.

Remarks. 1. It is commonplace in exploration seismology that seismic reflection data give much greater resolution of vertical rock structures than of horizontal variations, at least in the absence of substantial ray bending. As the present problem is a crude, straight-ray version of the seismic problem, the a priori constraint on horizontal derivatives of η^1, implied by $\eta^1 \in \mathfrak{R}$, seems quite natural.

2. Much sharper results are possible via propagation-of-singularities arguments—see [16] for such results for the (closely related) timelike Cauchy problem. In particular, the (regularized) dependence of η^1 on the data can be localized by ray-tracing. However, in that case the constants C depend on *higher* Sobolev norms of η^0 than appear in the above estimates for η^1; the dependence follows from results of Beals and Reed [2]. Such estimates force one to restrict the domain of $\mathscr{T}$ to a relatively compact subset to achieve continuity, whereas the inclusion $\mathscr{H}_{(1,s+2)} \subset H^1$ is not compact. Perhaps sharper propagation-of-regularity results would allow one to retain the local dependence property without such stringent regularization.

3. For layered media, i.e. η^0, η^1 independent of x, the inverse problem amounts to the much-studied impedance profile inversion problem, and the statement of Theorem 1 becomes (part of) the statement that $\mathscr{T}$ is a C^1-diffeomorphism $H^1[0,T] \to H^1[0,2T]$ in that case—see [15]. On the other hand, for topologies much weaker than H^1, $\mathscr{T}$ fails even to be continuous—see [6]—or $\mathscr{T}^{-1}$ fails to be continuous—see [1]. This accounts for our restriction to H^1 and related topologies. Also, rougher media may be approached via homogenization arguments.

4. Theorem 3 asserts uniqueness for the linearized inverse problem for the case $T = \infty$. For $n = 1$, a very similar argument yields uniqueness for $T < \infty$. We are prevented from making direct use of the one-dimensional pattern here by the lack of local uniqueness for the time-like Cauchy problem, and instead use a transform argument. The transform argument allows us to appeal to elliptic unique continuation results, but restricts us to the case $T = \infty$. Recently, Paul Sacks has used the full strength of the local part of the proof of Theorem 3, together with a very clever application of Hörmander's uniqueness theorem for the Cauchy problem, to extend Theorem 3 to the case $T < \infty$. Details may be found in [18].

5. The restriction of $\mathscr{A}$ in the statement of Theorem 4 is necessary because the sets $\mathfrak{R}(\gamma, \Gamma)$—or indeed their union over γ—are not weakly closed. An example of a function in the complement in $\mathscr{H}_{(1,s)}(\mathbb{R}^2_+)$ of $\cup_\gamma \mathfrak{R}(\gamma, \Gamma)$ is

$$\varphi(x) e^{-z} z^{s+1} \sin(xz^{-1})$$

with $\varphi \in C_0^\infty(\mathbb{R})$. Such functions may occur as the $\mathscr{H}_{(1,s)}$—weak limits of sequences in $\mathfrak{R}(\gamma, \Gamma)$. On the other hand, if we modify the definition of $\mathfrak{R}$ by requiring uniform integrability in x at infinity, rather than lateral compact support, then the class of uniformly x-bandlimited functions

$$\left\{ u \in \mathscr{H}_{(1,s+2)}(\mathbb{R}^n_+) : \hat{u}(z,\xi) = 0 \text{ for } |\xi| \geqq \Xi, z \geqq 0 \right\}$$

lies in the union of the modified $\mathfrak{R}(\gamma)$'s. Since Ξ is independent of γ in this regard and may be chosen as large as desired, perhaps the restriction is not particularly severe. For further discussion see [18], especially §5.

We shall need a number of the results of [12] concerning the relation between ρ^0 and u^0, as well as a technical lemma, which we now list:

(IV) ([12, Lemma 1]). Suppose $f \in \mathscr{S}(\mathbb{R}^n)$, $g_2 \in L^2(\mathbb{R}^n)$, $g_1 \in L^2(\mathbb{R}^n) \cap L^\infty(\mathbb{R}_z, L^2(\mathbb{R}_x^{n-1}))$, and $s > (n-1)/2$. Then for some $C = C(s,n) > 0$,

$$\left| \int_{\mathbb{R}} dz \int_{\mathbb{R}^{n-1}} dx f g_1 g_2 \right| \leqq C \|f\|_{(0,s)} \|g_2\|_{L^2(\mathbb{R}^n)} \sup_{z \in \mathbb{R}_z} \|g_1(\cdot, z)\|_{L^2(\mathbb{R}_x^{n-1})}. \tag{10}$$

Moreover, if f is independent of $x_1, \cdots, x_k$, then the same conclusion holds for $s \geqq \frac{1}{2}(n-k-1)$.

(V) An intermediate result in the proof of [12, Lemma 1] which we shall also need, is: suppose $f \in \mathscr{S}(\mathbb{R}^n)$, $g \in L^\infty(\mathbb{R}_z, L^1(\mathbb{R}_x^{n-1}))$. Then

$$\left| \int dz \int dx f g \right| \leqq C \int dz \|f(\cdot, z)\|_{H^s(\mathbb{R}_x^{n-1})} \|g(\cdot, z)\|_{L^1(\mathbb{R}_x^{n-1})} \tag{11}$$

with $C = C(s,n) > 0$.

Similarly,

$$\int dz \int dx f^2 |g| \leqq C \|f\|_{(0,s)}^2 \sup_z \|g(\cdot, z)\|_{L^1(\mathbb{R}^{n-1})}. \tag{12}$$

We also require the following Poincaré-style inequality:

(VI) Suppose that $s > (n-1)/2$, $\log a(\cdot, 0) \in H^{s+1}(\mathbb{R}^{n-1})$, $\log a \in \mathscr{H}_{(1,s+1)}(\mathbb{R}^{n-1} \times [0,T])$, $u(\cdot, 0) \in H^1(\mathbb{R}^{n-1})$, $u \in H^1(\mathbb{R}^{n-1} \times [0,T])$. Then for some constant $C > 0$ depending on $\|\log a(\cdot, 0)\|_s$, $\|\log a\|_{(1,s+1)}$, s, n, and T,

$$\|u\|_1^2 \leqq C \left\{ \|\nabla(au)\|_0^2 + \|u(\cdot, 0)\|_1^2 \right\}. \tag{13}$$

Proof of (13). First observe that $a_- \leqq a \leqq a_+$ with $a_\pm$ depending on $\|\log a(\cdot, 0)\|_S$, $\|\log a\|_{(1,s+1)}$, and T.

Next suppose $u(\cdot, 0) \equiv 0$. Then the simplest Poincaré inequality gives

$$\|au\|_1 \leqq C \|\nabla au\|_0. \tag{14}$$

Also

$$\|au(\cdot, z)\|_0 \leqq C \|au\|_1, \qquad 0 \leqq z \leqq T.$$

So

$$\begin{aligned} \|u\|_1^2 = \|u\|_0^2 + \|\nabla u\|_0^2 &\leqq a_-^{-1} \|au\|_0^2 + \|\nabla(a^{-1} \cdot au)\|_0^2 \\ &\leqq C \|au\|_0^2 + 2 \|(\nabla a^{-1}) au\|_0^2 + 2 a_-^{-1} \|\nabla au\|_0^2. \end{aligned} \tag{15}$$

But

$$\begin{aligned} \|(\nabla a^{-1}) au\|_0^2 &= \int_0^T dz \int dx (\nabla a^{-1})^2 |au|^2 \\ &\leqq \int_0^T dz \int dx |a^{-2} \nabla a|^2 \sup_z \|au(\cdot, z)\|_0^2 \leqq C \|au\|_1^2 \end{aligned}$$

which together with (15) gives

$$\|u\|_1^2 \leqq C \|au\|_1^2.$$

Combined with (14) this gives the result for the special case.

In general,

$$\begin{aligned}\|u\|_1 &\leqq \|u(\cdot,0)\|_1 + \|u-u(\cdot,0)\|_1 \\ &\leqq \|u(\cdot,0)\|_1 + C\|\nabla a(u-u(\cdot,0))\|_0 \\ &\leqq \|u(\cdot,0)\|_1 + C\{\|\nabla(au)\|_0 + \|\nabla(au(\cdot,0))\|_0\} \\ &\leqq \|u(\cdot,0)\|_1 + C\{\|\nabla au\|_0 + \|\nabla_x u(\cdot,0)\|_0 + \|u(\cdot,0)\|_0\}\end{aligned}$$

where in the last step we have used the estimate (12) with $f=\nabla a$ and $g=|u(\cdot,0)|^2$. Q.E.D.

(VII) The following estimates hold between u^0 and η^0 (each statement is followed by its reference in [12], of which it is (at most) a minor modification):

$$\int_0^{\min(T,t)} dz \int dx \eta^0\left(|\partial_t u^0|^2 + |\nabla u^0|^2\right)(x,z,t) := \|u^0(t)\|_E^2 \leqq \|\log \eta^0\|_1^2, \tag{16}$$

(Lemma 2),

$$\|D_j u^0(t)\|_E \leqq C\|\log \eta^0\|_{(1,s+1)} \qquad (j=1,\cdots,n-1), \tag{17}$$

(Lemma 3, Theorem 1),

$$\sup_{0\leqq z\leqq T}\int_z^{2T-z} dt \int dx\left(|\partial_t u^0|^2 + |\nabla u^0|^2\right)(x,z,t) \leqq CF\left(\|\log \eta^0\|_{(1,s+1)}, T\right), \tag{18}$$

(Theorem 1),

$$\|D_k D_j u^0(t)\|_E \leqq C\|\log \eta^0\|_{(1,s+2)} \qquad (k,j=1,\cdots,n-1), \tag{19}$$

(Lemma 5, Lemma 6),

$$\sup_{0\leqq z\leqq T}\int_z^{2T-z} dt \int dx\left\{|\partial_t D_k u^0|^2 + |\nabla D_k u^0|^2\right\}(x,z,t) \leqq CF\left(\|\log \eta^0\|_{(1,s+2)}, T\right), \tag{20}$$

(Lemma 6).

Here C denotes a constant depending on s, T, and n, and F is a smooth function of its arguments.

Proof of Theorem 1. We shall establish the first estimate. The second is proved by further arguments of the same sort, following the pattern of the proof of [15, Thm. 2.5]. Therefore we omit the proof of the second statement.

The main tool in the proof of *Theorem* 1 is the "sideways energy form"

$$Q(z) := \frac{1}{2}\int_z^{2T-z} dt \int dx\left\{|\partial_t u^1|^2 + |\nabla u^1|^2\right\}.$$

A short calculation yields

$$\begin{aligned}\frac{d}{dz}Q(z) = &-\int_z^{2T-z} dt \int dx\left\{\left(\partial_t^2 u^1 - \Delta u^1\right)\partial_z u^1 + 2\nabla_x u^1 \cdot \partial_z \nabla_x u^1\right\} \\ &-\frac{1}{2}\int dx\left[|\nabla \bar{u}^1|^2 + [\nabla \bar{\bar{u}}^1]^2\right](x,z)\end{aligned} \tag{21}$$

where $\bar{u}^1(x,z)=\bar{u}^1(x,z,z)$, as before, and $\bar{\bar{u}}(x,z) := u^1(x,z,2T-z)$.

Remark. The second term on the r.h.s. in (21) cannot be estimated in terms of Q alone, which reflects the nonhyperbolic nature of the timelike Cauchy problem.

Integrate (21) from z to T, make use of the boundary value problem (8) for u^1, and note that $Q(T)=0$ to obtain

$$
\begin{aligned}
Q(z) = -\int_z^T d\zeta \int_\zeta^{2T-\zeta} dt \int dx \Big\{ & \nabla \log \eta^0 \cdot \nabla u^1 \partial_z u^1 \\
& + \nabla \frac{\eta^1}{\eta^0} \cdot \nabla u^0 \partial_z u^1 + 2 \nabla_x u^1 \partial_z \nabla_x u^1 \Big\}(x,\zeta,t) \qquad (22) \\
& + \frac{1}{2} \int_z^T d\zeta \int dx \Big\{ |\nabla \bar{\bar{u}}^1|^2 + \frac{1}{4} \big| \nabla \big[\eta_0^{1/2} (\eta^0)^{-3/2} \eta^1 \big] \big|^2 \Big\}.
\end{aligned}
$$

Estimate the first term on the r.h.s. of (22) using (11): for $s > n/2 - 1$,

$$
\begin{aligned}
& \Big| \iiint \nabla \log \eta^0 \cdot \nabla \eta^1 \partial_z u^1 \Big| \\
& \quad \leqq C \int_z^T dz \| \nabla \log \eta^0(\cdot, z) \|_{H^s(\mathbb{R}^{n-1})} \| \nabla u^1 \partial_z u^1(\cdot, z, \cdot) \|_{L^1(\mathbb{R}^n)} \qquad (23) \\
& \quad \leqq C \int_z^T \sigma Q
\end{aligned}
$$

with

$$
\sigma(z) := \| \nabla \log \eta^0(\cdot, z) \|_{H^s(\mathbb{R}^{n-1})}.
$$

For the second term, use (10):

$$
\begin{aligned}
& \Big| \iiint \nabla \frac{\eta^1}{\eta^0} \cdot \nabla u^0 \partial_z u^1 \Big| \\
& \quad \leqq C \| \nabla \eta^1 \|_{(0,s)} \Big(\iiint |\partial_z u^1|^2 \Big)^{1/2} \sup_{z \leqq \zeta \leqq T} \| \nabla u^0(\cdot, \zeta, \cdot) \|_{L^2(\mathbb{R}^{n-1} \times [\zeta, 2T-\zeta])}.
\end{aligned}
$$

The first factor is bounded by $\|\eta^1\|_{(1,s+1)}$, the second by $(\int_z^T Q)^{1/2}$ and the third by $\|\eta^0\|_{(1,s+1)}$, according to (18). So we obtain

$$
\begin{aligned}
\Big| \iiint \nabla \frac{\eta^1}{\eta^0} \cdot \nabla u^0 \nabla_z u^1 \Big| & \leqq C \|\eta^1\|_{(1,s+1)} \Big(\int_z^T Q \Big)^{1/2} \\
& \leqq \frac{1}{2} C \Big(\|\eta^1\|_{(1,s+1)}^2 + \int_z^T Q \Big). \qquad (24)
\end{aligned}
$$

To estimate the third term, we require an energy estimate for the tangential derivatives $\nabla_x u^1$. Define $v_j = D_j u^1$, $j = 1, \cdots, n-1$. Then in $C_T := \{(x,z,t)\colon x \in \mathbb{R}^{n-1}, 0 \leqq z \leqq T, z \leqq t \leqq 2T - z\}$

$$
\begin{aligned}
& \big(\partial_t^2 - \nabla - \nabla \log \eta^0 \cdot \nabla \big) v_j = \nabla D_j \frac{\eta^1}{\eta^0} \cdot \nabla u^0 + \nabla \frac{\eta^1}{\eta^0} \cdot \nabla D_j u^0 + \nabla D_j \log \eta^0 \cdot \nabla u^1, \\
& \partial_z v_j(\cdot, 0, \cdot) = 0, \qquad (25) \\
& \bar{v}_j = -\frac{1}{2} D_j \big(\eta_0^{1/2} (\eta^0)^{-3/2} \eta^1 \big).
\end{aligned}
$$

Multiply the wave equation by $\eta^0 \partial_t v_j$ and integrate by parts over $\Omega \cap \{t \leqq \tau\}$ to obtain the energy identity

$$\iiint_{\Omega \cap \{t \leqq \tau\}} \eta^0 \left(\nabla D_j \frac{\eta^1}{\eta^0} \cdot \nabla u^0 + \nabla \frac{\eta^1}{\eta^0} \cdot \nabla D_j u^0 + \nabla D_j \log \eta^0 \cdot \nabla u^1 \right) \partial_t v_j$$

$$= -\frac{1}{2} \int_0^{\min(\tau, T)} dz \int dx \eta^0 |\nabla \bar{v}_j|^2 + \frac{1}{2} h(\tau - T) \int_{2T-\tau}^{T} dz \int dx \eta^0 |\nabla \bar{\bar{v}}_j|^2 \tag{26}$$

$$+ \frac{1}{2} \int_0^{\min(\tau, 2T-\tau)} dz \int dx \eta^0 \left\{ |\partial_t v_j|^2 + |\nabla v_j|^2 \right\}(x, z, \tau)$$

where h is the Heaviside function. Define $E_j(\tau)$ by the last term in (26). Estimate the first two terms on the l.h.s. of (26) by means of (11):

$$\left| \iiint_{\Omega \cap \{t \leqq \tau\}} \left(\nabla D_j \frac{\eta^1}{\eta^0} \cdot \nabla u^0 + \nabla \frac{\eta^1}{\eta^0} \cdot \nabla D_j u^0 \right) \partial_t v_j \right|$$

$$\leqq C \left\{ \|\eta^1\|_{(1, s+2)} \sup_{0 \leqq z \leqq T} \|\nabla u^0(\cdot, z, \cdot)\|_{L^2(\mathbb{R}^{n-1} \times [2T-z])} \right.$$

$$\left. + \|\eta^1\|_{(1, s+1)} \sup_{0 \leqq z \leqq T} \|\nabla D_j u^0(\cdot, z, \cdot)\|_{L^2(\mathbb{R}^{n-1} \times [z, 2T-z])} \right\} \left(\iiint_{\Omega \cap \{t \leqq T\}} |\partial_t v_j|^2 \right)^{1/2}.$$

The suprema in the above estimate are bounded in terms of $\|\eta^0\|_{(1, s+2)}$ according to (18) and (20). Also, the last factor is the integrated energy. Thus the left-hand side above is bounded by

$$\leqq C \left(\|\eta^1\|_{(1, s+2)}^2 \int_0^\tau E_j \right)^{1/2}$$

$$\leqq \frac{1}{2} C \left\{ \|\eta^1\|_{(1, s+2)}^2 + \int_0^\tau E_j \right\}. \tag{27}$$

Also,

$$\iiint_{\Omega \cap \{t \leqq \tau\}} |\nabla D_j \log \eta^0 \cdot \nabla u^1 \partial_t v_j|$$

$$\leqq C \sup_{0 \leqq z \leqq T} \|\nabla u^1(\cdot, z, \cdot)\|_{L^2(\mathbb{R}^{n-1} \times [z, 2T-z])} \times \left(\iiint_{\Omega \cap \{t \leqq \tau\}} |\partial_t v_j|^2 \right)^{1/2}.$$

Define $Q^* = \sup_{0 \leqq z \leqq T} Q(z)$. Then the above is

$$\leqq C \left(Q^* \int_0^\tau E_j \right)^{1/2}$$

$$\leqq \frac{1}{2} C \left(Q^* + \int_0^\tau E_j \right). \tag{28}$$

Combine (27), (28), and a slightly rearranged version of (26) to obtain

$$E_j(\tau) \leqq C \left\{ \|\eta^1\|_{(1, s+2)}^2 + Q^* + \int_0^\tau E_j \right\}$$

whence Gronwall's lemma implies

$$E_j(\tau) \leqq Ce^{C\tau}\left\{\|\eta^1\|^2_{(1,s+2)} + Q^*\right\}.$$

Now

$$\int_z^T d\zeta \int_\zeta^{2T-\zeta} dt \int dx \left|\nabla \cdot D_j u^2\right|^2 \leqq \int_0^{2T} E_j \leqq C\left(\|\eta^1\|^2_{(1,s+2)} + Q^*\right). \tag{29}$$

We also need an energy estimate for u^1. Multiply the wave equation in (8) by $\eta^0 \partial_t u^1$ and integrate by parts over $\Omega \cap \{t \leqq \tau\}$ to obtain

(30)

$$\begin{aligned}\iiint_{\Omega\cap\{t\leqq\tau\}} \left(\nabla \frac{\eta^1}{\eta^0}\cdot \nabla u^0\right)\partial_t u^1 = &-\frac{1}{2}\int_0^{\min(\tau,T)} dz \int dx\, \eta^0 |\nabla \bar{u}^1|^2 \\ &+ \frac{1}{2} h(\tau - T)\int_{2T-\tau}^T dz \int dx\, \eta^0 |\nabla \bar{\bar{u}}^1|^2 \\ &+ \frac{1}{2}\int_0^{\min(\tau,2T-\tau)} dz \int dx\, \eta^0 \left\{|\partial_t u^1|^2 + |\partial u^1|^2\right\}(x,z,\tau).\end{aligned}$$

Define E to be the last term above, and estimate the l.h.s. as before to obtain

$$E(\tau) \leqq C\left\{\|\eta^1\|^2_{(1,s+1)} + \int_0^\tau E\right\}$$

whence

$$E(\tau) \leqq Ce^{Ct}\|\eta^1\|^2_{(1,s+1)}. \tag{31}$$

Now estimate the third term in (22):

$$\begin{aligned}&\left|\int_z^T d\zeta \int_\zeta^{2T-\zeta} dz \int dx\left(\nabla_x u^1 \cdot \nabla_z \nabla_x u^1\right)\right| \\ &\leqq \left(\iiint |\nabla_x u^1|^2 \iiint |\partial_z \nabla_x u^1|^2\right)^{1/2} \\ &\leqq \left(\int_0^T E\right)^{1/2}\left(\sum_{j=1}^{n-1}\iiint \left|\left(\nabla \cdot D_j u^1\right)\right|^2\right)^{1/2} \\ &\leqq C\|\eta^1\|_{(1,s+1)}\left(\|\eta^1\|^2_{(1,s+2)} + Q^*\right)^{1/2}.\end{aligned} \tag{32}$$

To estimate the fourth term, note that (30) for $\tau = 2T$ reads

$$\begin{aligned}&\frac{1}{2}\int_0^T dz \int dx\, \eta^0 |\nabla \bar{\bar{u}}^1|^2 \\ &= \frac{1}{2}\int_0^T dz \int dx\, \eta^0 |\nabla \bar{u}^1|^2 + \iiint_\Omega \left(\nabla \frac{\eta^1}{\eta^0}\cdot \nabla u^0\right)\partial_t u^1 \\ &\leqq C\left\{\|\eta^1\|_1^2 + \|\eta^1\|_{(1,s+1)} \sup \|\nabla u^0(\cdot,z,\cdot)\|_{L^2(\mathbb{R}^{n-1}\times[z,2T-z])}\left(\iiint_\Omega |\partial_t u^1|^2\right)^{1/2}\right\} \\ &\leqq C\|\eta^1\|^2_{(1,s+1)}\end{aligned} \tag{33}$$

using bounds quoted above. Since η^0 is bounded below uniformly by a function of $\|\log\eta^0\|_{(1,s+1)}$, we have estimated the fourth term. In the process, we have estimated the fifth term also:

$$\int_0^T dz \int dx|\nabla\bar{u}^1|^2 \leqq C\|\eta^1\|_1^2. \tag{34}$$

Now, combining (22), (23), (24), (32), (33), and (34), we have

$$Q(z) \leqq C\left\{\|\eta^1\|_{(1,s+2)}^2 + \|\eta^1\|_{(1,s+1)}\left(\|\eta^1\|_{(1,s+2)}^2 + Q^*\right)^{1/2} + \int_z^T (\sigma+1)Q\right\}.$$

Now Gronwall's inequality implies

$$Q(z) \leqq C\|\eta^1\|_{(1,s+2)}\left(1+\left(\|\eta^1\|_{1,s+2}^2 + Q^*\right)^{1/2}\right), \qquad 0 \leqq z \leqq T.$$

Since we can replace $Q(z)$ by Q^* on the l.h.s., it follows immediately that Q is bounded by a function of $\|\eta^1\|_{1,s+2}$ and of course, of n, s, T and $\|\log\eta^0\|_{(1,s+2)}$, which finishes the proof of Theorem 1. For use in the proof of Theorem 3, note that inequality (i) together with (32) implies

$$E_j(\tau) \leqq C_1 e^{C_2 t}\|\eta^1\|_{(1,s+2)}^2. \tag{35}$$

Q.E.D.

For use in the proof of Theorem 2, we note another obvious consequence:

COROLLARY 1. (Same hypotheses as Theorem 1.) *For* $0 \leqq z \leqq T$,

$$\|\nabla_x u^1(\cdot, z, \cdot)\|_{L^2(\mathbb{R}^{n-1}\times[z,2T-z])} \leqq C\|\eta^1\|_{(1,s+2)}.$$

Proof. Integrating (35) from $\tau=0$ to $\tau=2T$, we see that $\nabla_x u^1$ has Dirichlet norm in C_T bounded by $C\|\eta^1\|_{(1,s+2)}$. By the transport equation its boundary value on $z=t$ is bounded by a similar quantity, so its L^2 norm, thus its H^1 norm, in C_T is also bounded by $C\|\eta^1\|_{(1,s+2)}$. Now the standard trace theorem implies the result. Q.E.D.

Proof of Theorem 2. The first step is to devise an identity similar to (22), involving an indefinite form: set

$$P(z) = \frac{1}{2}\int_z^{2T-z} dt \int dx\left\{\left(\partial_t u^1\right)^2 + \left(\partial_z u^1\right)^2 - \left|\nabla_x u^1\right|^2\right\}.$$

Then a short calculation yields

$$\begin{aligned}\partial_z P(z) = &-\frac{1}{2}\int dx\left\{\left[\left(\partial_z\bar{u}^1\right)^2 + \left(\partial_z\bar{\bar{u}}^1\right)^2\right] - \left[\left(\nabla_x\bar{u}^1\right)^2 + \left(\nabla_x\bar{\bar{u}}^1\right)^2\right]\right\} \\ &- \int_z^{2T-z} dt \int dx\left\{\nabla\log\eta^0\cdot\nabla u^1 + \nabla\left(\frac{\eta^1}{\eta^0}\right)\cdot\nabla u^0\right\}\partial_z u^1.\end{aligned} \tag{36}$$

Now

$$P(z) = Q(z) - 2\int_z^{2T-z} dt \int dx\left|\nabla_x u^1\right|^2.$$

Therefore, integration of (36) from $z=0$ to $z=T$ and some manipulation yields

$$
\begin{aligned}
&Q(z)+\frac{1}{2}\int_0^z d\zeta\int dx\left[|\nabla\bar{u}^1|^2+|\nabla\bar{\bar{u}}^1|^2\right]+\frac{1}{2}\|\eta^1(\cdot,0)\|_1^2\\
&\quad=Q(0)+2\int_z^{2T-z}dt\int dx|\nabla_x u^1|^2(x,z,t)-2\int_0^{2T}dt\int dx|\nabla_x u^1|^2(x,0,t)\\
&\quad\quad+\int_0^z d\zeta\int dx\left[|\nabla_x\bar{u}^1|^2+|\nabla_x\bar{\bar{u}}^1|^2\right]\\
&\quad\quad-\int_0^z dz\int_z^{2T-z}dt\int dx\left\{\nabla\log\eta^0\cdot\nabla\left(\frac{\eta^1}{\eta^0}\right)\cdot\nabla u^0\right\}\partial_z u^1+\frac{1}{2}\|\eta^1(\cdot,0)\|_1^2.
\end{aligned}
\tag{37}
$$

Recall that

$$
\bar{u}^1=-\frac{1}{2}\eta_0^{1/2}(\eta^0)^{-3/2}\eta^1.
$$

With $u\to\eta^1$ and $a\to\frac{1}{2}\eta_0^{1/2}(\eta^0)^{-3/2}$, the Poincaré inequality (13) implies the existence of a constant $C>0$, depending on η_0, η^0, and T, for which

$$
C\|\eta^1\|^2_{H^1(\mathbb{R}^{n-1}\times[0,z])}\leqq\frac{1}{2}\int_0^z d\zeta\int dx|\nabla\bar{u}^1|^2+\frac{1}{2}\|\eta^1(\cdot,0)\|_1^2.
\tag{38}
$$

Now (37) and (38) combine to yield for any $\alpha>0$,

$$
\begin{aligned}
&Q(z)+C_1\|\eta^1\|^2_{H^1(\mathbb{R}^{n-1}\times[0,z])}\\
&\quad\leqq Q(0)+2\|\nabla_x u^1(\cdot,z,\cdot)\|_0^2-2\|\nabla_x u^1(\cdot,0,\cdot)\|_0^2\\
&\quad\quad+\|\nabla_x\bar{u}^1\|^2_{L^2(\mathbb{R}^{n-1}\times[0,z])}+\|\nabla_x\bar{\bar{u}}^1\|^2_{L^2(\mathbb{R}^{n-1}\times[0,z])}\\
&\quad\quad+\int_0^z\left(\sigma+C_2\alpha^{-1}\right)Q+\alpha\left\|\nabla\frac{\eta^1}{\eta^0}\right\|^2_{\mathscr{H}_{(0,s)}(\mathbb{R}^{n-1}\times[0,z])}+\frac{1}{2}\|\eta^1(\cdot,0)\|_1^2.
\end{aligned}
\tag{39}
$$

Since

$$
\left\|\nabla\frac{\eta^1}{\eta^0}\right\|_{\mathscr{H}_{(0,s)}(\mathbb{R}^{n-1}\times[0,z])}\leqq C\|\eta^1\|_{\mathscr{H}_{(1,s+1)}(\mathbb{R}^{n-1}\times[0,z])}\leqq C\|\eta^1\|_{H^1(\mathbb{R}^{n-1}\times[0,z])}
$$

(the last inequality by virtue of $\eta^1\in\mathfrak{R}(\gamma,\Gamma)$), if we choose α small enough we can absorb the next-to-the-last term in (39) into the l.h.s., at the price of decreasing C_1 somewhat. Now Gronwall's inequality implies

$$
Q(z)+C_1\|\eta^1\|^2_{H^1(\mathbb{R}^{n-1}\times[0,z])}\leqq\left\{Q(0)+K(\eta^1)\right\}\exp\int_0^z\left(\sigma+C_2\alpha^{-1}\right)
$$

so ($Q(T)=0$)

$$
\|\eta^1\|_1^2\leqq C\left\{Q(0)+K(\eta^1)\right\}
$$

where we have defined

$$K(\eta^1) := -2\|\nabla_x u^1(\cdot,0,\cdot)\|^2 + \sup_{0\leqq z\leqq T}\Big\{\|\nabla_x \bar{u}^1\|^2_{L^2(\mathbb{R}^{n-1}\times[0,z])} + \|\nabla_x \bar{\bar{u}}^1\|^2_{L^2(\mathbb{R}^{n-1}\times[0,z])} + 2\|\nabla_x u^1(\cdot,z,\cdot)\|_{L^2(\mathbb{R}^{n-1}\times[z,2T-z])}\Big\} + \frac{1}{2}\|\eta^1(\cdot,0)\|_1^2.$$

It remains to verify the property claimed for K. Accordingly suppose $\{\eta_j^1\}\subset \mathfrak{R}(\gamma,\Gamma)$ with $\eta_j^1 \to 0$ weakly in $H^1(\mathbb{R}^{n-1}\times[0,T])$. Then $\{\eta_j^1\}$ is bounded in $\mathscr{H}_{(1,s+2)}(\mathbb{R}^{n-1}\times[0,T])$ and so has a subsequence converging weakly in $\mathscr{H}_{(1,s+2)}$. Since the (weak) limit must be the same distribution as the (weak) limit in H^1, the limit is unique ($=0$); therefore the entire sequence tends weakly to zero in $\mathscr{H}_{(1,s+2)}$. Since $\{\eta_j^1\}$ share the same compact support, the supports of the u_j^1 are also contained in the same compact subset of C_T, and likewise for $\operatorname{supp}\nabla_x u^1(\cdot,z,\cdot)$ for each $z\in[0,T]$. Now the trace and embedding theorems for Sobolev classes (e.g. [7, Thms. 2.5.6 and 2.2.3]) imply that maps

$$\eta^1 \to \nabla_x \bar{u}^1,$$
$$\eta^1 \to \nabla_x \bar{\bar{u}}^1,$$
$$\eta^1 \to \eta^1(\cdot,0),$$
$$\eta^1 \to \nabla_x u^1(\cdot,z,\cdot)$$

from $\mathscr{H}_{(1,s+2)}(\Gamma) := \{f\in\mathscr{H}_{(1,s+2)}\mathbb{R}^{n-1}\times[0,T]\colon \operatorname{supp} f\subset\Gamma\}$ to $L^2(\mathbb{R}^{n-1}\times[0,z])$, $L^2(\mathbb{R}^{n-1}\times[0,z])$, $H^1(\mathbb{R}^{n-1})$, and $L^2(\mathbb{R}^{n-1}\times[z,2T-z])$ respectively, are *compact* for each $z\in[0,T]$. We conclude that

$$\|\nabla_x u_j^1(\cdot,0,\cdot)\|_{L^2(\mathbb{R}^{n-1}\times[0,2T])}\to 0, \qquad j\to\infty,$$

and for each z,

$$F_j(z) := \|\nabla_x \bar{u}_j^1\|^2_{L^2(\mathbb{R}^{n-1}\times[0,z])} + \|\nabla_x \bar{\bar{u}}_j^1\|^2_{L^2(\mathbb{R}^{n-1}\times[0,z])} + 2\|\nabla_x u_j^1(\cdot,z,\cdot)\|^2_{L^2(\mathbb{R}^{n-1}\times[z,2T-z])}\to 0, \qquad j\to\infty.$$

We claim that $\{F_j\}$ is uniformly Hölder-continuous with exponent one-half. In fact,

$$\partial_z F_j(z) = 2\int_z^{2T-z} dt\int dx\, \nabla_x u_j^i\cdot\nabla_x\partial_z u_j^i$$

so

$$|F_j(z_1)-F_j(z_2)|\leqq 2\|\nabla_x u_j^1\|\,\|\nabla_x\partial_z u_j^1\|$$

(where the norms are $L^2(C_T\cap\{z_1\leqq z\leqq z_2\})$). But $\|\nabla_x u_j^1\|_{H^1(C_T)}$ is bounded by $C\|\eta_j^1\|_{(1,s+2)}$; hence uniformly in j, as is $\|\nabla_x u_j^1(\cdot,z,\cdot)\|_{L^2(\mathbb{R}^{n-1}\times[z,2T-z])}$, $0\leqq z\leqq T$ (Corollary 1), and

$$\|\nabla_x u_j^1\|\leqq\sqrt{(z_2-z_2)}\sup_{z_1\leqq z\leqq z_2}\|\nabla_x u_j^1(\cdot,z,\cdot)\|_{L^2(\mathbb{R}^{n-1}\times[z,2T-z])}$$

so

$$|F_j(z_1)-F_j(z_2)| \leqq C|z_1-z_2|^{1/2}$$

with C independent of j.

Now let $\varepsilon>0$, and let $\{z_k\}_{k=0}^N$ be an ordered partition of $[0,T]$ with

$$\max(z_k-z_{k-1})<\varepsilon^2.$$

By the pointwise result derived earlier, for some $J\in\mathbb{N}$, $j\geqq J$ implies

$$F_j(z_k)<\varepsilon, \qquad k=0,\cdots,N.$$

On the other hand, if $z\in[z_k,z_{k+1})$,

$$|F_j(z)|\leqq|F_j(z_k)|+|F_j(z_k)-F_j(z)|\leqq(1+C)\varepsilon.$$

Since ε is arbitrary, we have established that $F_j\to 0$ uniformly in $[0,T]$. Since

$$K(\eta_j^1)=-2\|\nabla_x u^1(\cdot,0,\cdot)\|_0^2+\frac{1}{2}\|\eta^1(\cdot,0)\|_1+\|F_j\|_{CU^0[0,T]},$$

the theorem is proved. Q.E.D.

Proof of Theorem 3. We begin by establishing a local uniqueness result.

The hypotheses imply that, for any $T>0$, $Q(0)=0$. Now integrate (21) from 0 to z to obtain

$$\begin{aligned} &Q(z)+\frac{1}{2}\int_0^z d\zeta\int dx|\nabla\bar{u}^1|^2(x,\zeta)\\ &\leqq\int_0^z d\zeta\int_\zeta^{2T-\zeta}dt\int dx\left\{2|\nabla_x u^1\cdot\nabla_x\partial_z u^1|\right.\\ &\qquad\left.+|\nabla\log\eta^0\cdot\nabla u^1\partial_z u^1|+\left|\nabla\frac{\eta^1}{\eta^0}\cdot\nabla u^0\partial_z u^1\right|\right\}. \end{aligned}\tag{40}$$

From (35),

$$\iiint_\Omega|\nabla_x\partial_z u^1|^2\leqq C\|\eta^1\|_{(1,s+2)}^2.$$

(For the moment, $\|\eta^1\|_{(1,s+2)}=\|\eta^1\|_{\mathscr{H}_{(1,s+2)}(\mathbb{R}^{n-1}\times[0,T])}$, etc.). Therefore, for any $\alpha>0$,

$$\begin{aligned} &\int_0^z d\zeta\int_\zeta^{2T-\zeta}dt\int dx|\nabla_x u^1\cdot\nabla_x\partial_z u^1|\\ &\leqq C\|\eta^1\|_{(1,s+2)}\left(\int_0^z d\zeta\int_\zeta^{2T-\zeta}dt\int dx|\nabla_x u^1|^2\right)^{1/2}\\ &\leqq C\left(\alpha\|\eta^1\|_{(1,s+2)}^2+\alpha^{-1}\int_0^z Q\right). \end{aligned}\tag{41}$$

The second term is estimated as in (23):

$$\int_0^z dz\int_\zeta^{2T-\zeta}dt\int dx|\nabla\log\eta^0\cdot\nabla u^1\cdot\partial_z u^1|\leqq C\int_0^z\sigma Q.\tag{42}$$

The third term is estimated as in (24):

$$\int_0^z d\zeta \int_\zeta^{2T-\zeta} dt \int dx \left| \nabla \frac{\eta^1}{\eta^0} \cdot \nabla u^0 \partial_z u^1 \right| \leqq C\left(\alpha \|\eta^1\|^2_{(1,s+2)} + \alpha^{-1} \int_0^z Q \right). \tag{43}$$

The upshot of (40), (41), (42), and (43) is

$$Q(z) + \frac{1}{2} \int_0^z d\zeta \int dx \left| \nabla \left[\eta_0^{1/2} (\eta^0)^{-3/2} \eta^1 \right] \right|^2 \leqq C\left\{ \alpha \|\eta^1\|^2_{(1,s+2)} + \int_0^z (\sigma + \alpha^{-1}) Q \right\}.$$

Consequently,

$$Q(z) + \frac{1}{2} \int_0^z d\zeta \int dx \left| \nabla \left[\eta_0^{1/2} (\eta^0)^{-3/2} \eta^1 \right] \right|^2 \leqq C\alpha \|\eta^1\|^2_{(1,s+2)} \exp\left[C \int_0^z (\sigma + \alpha^{-1}) \right].$$

Estimate the Dirichlet integral on the l.h.s. from below by means of the Poincaré inequality (13), as in the proof of the previous theorem, and replace $\|\eta^1\|_{(1,s+2)} = \|\eta^1\|_{\mathscr{H}_{(1,s+2)}(\mathbb{R}^{n-1}\times[0,T])}$ by $C_0\|\eta^1\|_1$, as we can since $\eta^1 \in \mathfrak{R}(\gamma, \mathbb{R}^n_+)$. The upshot is

$$C_1 \|\eta^1\|_1^2 \leqq e^{C_2 T/\alpha} \exp\left[C_2 \int_0^T \sigma \right] C_3 \alpha \|\eta^1\|_1^2. \tag{44}$$

If α is sufficiently small, and $T = 0(\alpha)$, then

$$e^{C_2 T/\alpha} \exp\left[C_2 \int_0^T \sigma \right] C_3 \alpha < C_1, \tag{45}$$

whence $\|\eta^1\|_1 = 0$ from (44), i.e.: $\eta^1 \equiv 0$ in $\mathbb{R}^{n-1} \times [0, T]$ for some (small) $T > 0$.

Now we shall show that, in fact, $u^1 \equiv 0$ in $\mathbb{R}^{n-1} \times [0, T] \times \mathbb{R}^+$. It will follow that the Cauchy data for u^1 vanish on $\{z = T\}$, so we can repeat the above argument to show that η^1 vanishes on a somewhat larger interval. The constants in (45) depend only on $\|\log \eta^0\|_{(1,s+2)}$, s, and n, and T, are smooth functions of T near $T = 0$. Therefore, the thickness δT of the slab $\{T \leqq z \leqq T + \delta T\}$ in which we conclude $\eta^1 \equiv 0$, provided $\eta^1 \equiv 0$ in $\{0 \leqq z \leqq T\}$, is bounded below independently of T. Thus for any $T > 0$, finitely many repetitions of this "continuation" procedure yield $\eta^1 \equiv 0$ on $\{0 \leqq z \leqq T\}$, and so we may conclude $\eta^1 \equiv 0$, which finishes the proof modulo proof of the claim.

Remark. For $n = 1$, this procedure yields uniqueness up to $z = T$ where Cauchy data for u^1 on $\{z = 0\}$ are required to vanish only on $\{0 \leqq t \leqq 2T\}$ (see [15, proof of Theorem 2.8]), because of local uniqueness in the time-like Cauchy problem for $n = 1$. Local uniqueness in the time-like Cauchy problem for the wave operator modulo lower-order terms no longer holds for $n > 1$, as a consequence of a theorem of Hörmander [7, Thm. 8.9.4]. Consequently we must require that the Cauchy data vanish globally in order to carry out the "continuation" step.

To show that $u^1 \equiv 0$ for $0 \leqq z \leqq T$, recall the estimates (31) and (35) for the energy of u^1 and $\nabla_x u^1$:

$$E(t) \leqq C_1 e^{C_2 t} \|\eta^1\|^2_{(1,s+1)},$$
$$E_j(t) \leqq C_3 e^{C_2 t} \|\eta^1\|^2_{(1,s+2)}, \qquad j = 1, \cdots, n.$$

Since $u^1 \equiv 0$ on $z = 0$, it actually follows that

$$\|u^1(\cdot,\cdot,t)\|^2_{H^1(\mathbb{R}^{n-1}\times[0,t])} + \|\partial_t u^1(\cdot,\cdot,t)\|^2_{L^2(\mathbb{R}^{n-1}\times[0,t])}$$
$$+ \|\nabla_x u^1(\cdot,\cdot,t)\|^2_{H^1(\mathbb{R}^{n-1}\times[0,t])} + \|\partial_t \nabla_x u^1(\cdot,\cdot,t)\|^2_{L^2(\mathbb{R}^{n-1}\times[0,t])}$$

grows exponentially in t. Note that the proof of the local result above actually shows that $u^1 \equiv 0$ in $\{0 \leqq z \leqq T,\ z \leqq t \leqq 2T - z\}$. Thus $u^1(\cdot,\cdot,t)$ is actually of class H^1 in the full cylinder $\mathbb{R}^{n-1}\times[0,T]$ for $0 \leqq t \leqq \infty$, as is $\nabla_x u^1$. Furthermore, the local H^1 norms in $\mathbb{R}^{n-1}\times[0,t]$ a fortiori grow at exponential rate C_2. Therefore the Laplace transform $\tilde{u}^1(x,z,s)$ exists for $s > C_2$, and satisfies

$$\tilde{u}^1(\cdot,\cdot,s) \in H^1(\mathbb{R}^{n-1}\times[0,t]), \tag{46a}$$

$$\nabla_x \tilde{u}^1(\cdot,\cdot,s) \in H^1(\mathbb{R}^{n-1}\times[0,t]), \tag{46b}$$

$$(s^2 - \nabla^2 - \nabla \log \eta^0 \cdot \nabla)\tilde{u}^1 \equiv 0, \quad \text{in } \mathbb{R}^{n-1}\times[0,t], \tag{46c}$$

$$\tilde{u}^1 = \partial_z \tilde{u}^1 \equiv 0 \quad \text{on } \{z = 0\}, \quad C_2 < s < \infty. \tag{46d}$$

Since we have assumed $\nabla \log \eta^0 \in L^\infty(\mathbb{R}^n_+)$, it follows from (46b) and (46c) that $\partial_z \tilde{u}^1 \in H^1(\mathbb{R}^{n-1}\times[0,T])$; hence $\tilde{u}^1 \in H^2(\mathbb{R}^{n-1}\times[0,t])$. Therefore, we can extend $\tilde{u}^1$ by zero to negative z with $\tilde{u}^1$ remaining in H^2_{loc}. We are now in position to apply [7, Thm. 8.9]. It is easy to check that any plane is strongly pseudoconvex with respect to the principal part $-\nabla^2$ of the operator on the l.h.s. of (46c). Therefore, $\tilde{u}^1$ vanishes in $0 \leqq z \leqq T$, and by uniqueness of the Laplace transform, so does u^1.

As noted above, this finishes the proof of Theorem 3. Q.E.D.

Proof of Theorem 4. The assertion of the theorem is equivalent to the following:

Suppose $\log \eta^0 \in \mathscr{H}_{(1,s+2)}(\mathbb{R}^n_+) \cap W^{1,\infty}_{\text{loc}}(\mathbb{R}^n_+)$, $\{\eta^1_j\} \subset \mathscr{M}_\gamma$, $\eta^1_j(\cdot,0) \to 0$ in $H^1(\mathbb{R}^{n-1})$, and $D\mathscr{T}_\infty(\eta^0)\cdot\eta^1_j \to 0$ in $H^1_{\text{loc}}(\mathbb{R}^n_+)$. Then $\eta^1_j \to 0$ in $H^1_{\text{loc}}(\mathbb{R}^n_+)$ (hence in $\mathscr{H}_{(1,s+2)_{\text{loc}}}(\mathbb{R}^n_+)$).

Since $\mathfrak{R}(\gamma,\Gamma) \subset H^1(\mathbb{R}^n_1)$, it does no harm to assume $\{\eta^1_j\}$ bounded in H^1, hence (passing to a subsequence) weakly convergent in H^1_{loc}. Since $\mathfrak{R}(\gamma,\Gamma) \subset \mathscr{H}_{(1,s+2)}(\mathbb{R}^n_+)$ also, we can as well assume $\{\eta^1_j\}$ bounded and weakly convergent in $\mathscr{H}_{(1,s+2),\text{loc}}$. Since $\mathscr{H}_{(1,s+2)}$ is stronger than H^1_{loc}, the weak limit is unique; denote by η^1_∞ the weak limit. Now Theorem 1 implies

$$D\mathscr{T}_T(\eta^0)\eta^1_\infty = w - \lim D\mathscr{T}_T(\eta^0)\eta^1_j = 0$$

for any $T > 0$, so $D\mathscr{T}_\infty(\eta^0)\eta^1_\infty = 0$. Also, the hypothesis on $\mathscr{M}$ implies that $\eta^1_\infty \in \mathscr{M}$; hence $\eta^1_\infty \in \mathfrak{R}(\gamma',\Gamma)$ for some $\gamma' > 0$. It follows immediately from Theorem 3 that $\eta^1_\infty = 0$. Since $\{\eta^1_j\} \to 0$ weakly in $H^1(\mathbb{R}^n_+)$, a fortiori $\{\eta^1_j\} \to 0$ weakly in $H^1(\mathbb{R}^{n-1}\times[0,T])$ for each $T > 0$. Also, since $\Gamma \cap (\mathbb{R}^{n-1}\times[0,T])$ is compact, and thus $\text{supp}(D\mathscr{T}_\infty(\eta^0)\cdot\eta^1_j) \cap (\mathbb{R}^{n-1}\times[0,T])$ is compact, uniformly in j, convergence to zero of $D\mathscr{T}_\infty(\eta^0)\cdot\eta^1_j$ in H^1_{loc} implies convergence to zero of $D\mathscr{T}_T(\eta^0)\cdot\eta^1_j$ in $H^1(\mathbb{R}^{n-1}\times[0,T])$. We conclude from Theorem 2 that $\eta^1_j \to 0$ strongly in $H^1(\mathbb{R}^{n-1}\times[0,T])$. Since T is arbitrary, the proof is complete. Q.E.D.

Remark. A number of Soviet mathematicians have derived uniqueness and continuous dependence results for inverse problems of hyperbolic partial differential equations, under the assumptions that the coefficients may be expressed as finite sums

$$\eta(x,z) = \sum a_i(z)\varphi_i(x)$$

the φ_i being some prescribed (smooth) functions of x. (See, for instance, V. G. Jahno, *Uniqueness theorem for an inverse problem for a hyperbolic partial differential equation*, Differential Equations, 13 (1977), pp. 544–551.) Note that the coefficients in this class belong to the regularizing set $\mathfrak{R}(\gamma, \Gamma)$ for suitable γ, Γ. Thus, these results seem to be related to those presented here.

Acknowledgments. I would like to thank Paul Sacks and Percy Deift for useful conversations leading to several corrections and revisions. I would also like to thank the referee for pointing out an error in an example.

REFERENCES

[1] A. BAMBERGER, G. CHAVENT, AND P. LAILLY, *About the stability of the inverse problem in 1-D wave equations*, Appl. Math. Opt., 5 (1979), pp. 1–47.

[2] M. BEALS AND M. REED, *Propagation of singularities for hyperbolic pseudodifferential operators with nonsmooth coefficients*, Comm. Pure Appl. Math., 35 (1982), pp. 189–204.

[3] J. COHEN, N. BLEISTEIN, AND M. LAHLOU, *Highly accurate inversion methods for three-dimensional stratified media*, SIAM J. Appl. Math., 43 (1983), pp. 726–758.

[4] R. COURANT AND D. HILBERT, *Methods of Mathematical Physics* II, J. Wiley/Interscience, New York, 1962.

[5] F. G. FRIEDLANDER, *Sound Pulses*, Oxford Univ. Press, Cambridge, 1958.

[6] M. GERVER, *Inverse problem for the one-dimensional wave equation*, Geophys. J. Roy. Astr. Soc., 21 (1970), pp. 337–357.

[7] L. HÖRMANDER, *Linear Partial Differential Operators*, Springer-Verlag, New York, Berlin, Heidelberg, 1964.

[8] M. LAVRIENTIEV, V. ROMANOV, AND V. VASILIEV, *Multidimensional Inverse Problems for Differential Equations*, Lecture Notes in Mathematics 167, Springer-Verlag, New York, Berlin, Heidelberg, 1970.

[9] C. MORAWETZ, *A formulation for higher dimensional inverse problems for the wave equation*, Comp. Maths. Appl., 7 (1981), pp. 319–331.

[10] C. MORAWETZ AND G. KRIEGSMANN, *The calculations of an inverse potential problem*, SIAM J. Appl. Math., 43 (1983), pp. 844–854.

[11] R. STOLT AND A. JACOBS, *An approach to the inverse seismic problems*, preprint, 1981.

[12] W. SYMES, *Some aspects of inverse problems in several-dimensional wave propagation*, Proc. Conference on Inverse Problems, SIAM–AMS Proceedings 14, American Mathematical Society, Providence, RI, 1983.

[13] _____, *Impedance profile inversion via the first transport equation*, J. Math. Anal. Appl., 94 (1983), pp. 435–453.

[14] _____, *A trace theorem for solutions of the wave equation and the remote determination of acoustic sources*, Math. Math. Sci., 5 (1983).

[15] _____, *On the relation between coefficient and boundary values for solutions of Webster's horn equation*, this Journal, to appear.

[16] _____, *Continuation for solutions of wave equations: regularization of the time-like Cauchy problem*, preprint, 1985.

[17] J. RALSTON, *Gaussian beams and the propagation of singularities*, in Studies in Partial Differential Equations, W. Littman, ed., Mathematical Association of America, 1982.

[18] P. SACKS AND W. SYMES, *Uniqueness and continuous dependence for a multidimensional hyperbolic inverse problem*, Comm. PDE, to apear.

SIAM J. MATH. ANAL.
Vol. 17, No. 1, January 1986

A NONLINEAR INTEGRAL OPERATOR ARISING FROM A MODEL IN POPULATION GENETICS IV. CLINES*

ROGER LUI†

Abstract. We study the existence, uniqueness and stability properties of solutions to the integral equation $\phi = Q[\phi]$ with $\phi(-\infty)=1$, $\phi(\infty)=0$. Here $Q[u](x)=\int K(x-y)g(y,u(y))\,dy$ is defined on functions bounded between 0 and 1, K is a probability density function and $g(x,u)=[s(x)u^2+u]/[1+s(x)u^2+\sigma(x)(1-u)^2]$ according to a population genetics model. The hypotheses on g are based on the biological assumption that the homozygotes, that is individuals with genotypes AA or aa, are best fit to survive near opposite ends of the one-dimensional habitat.

1. Introduction. In the first section of [13] a population genetics model was formulated that describes the change in gene fractions over successive generations of a population living in a homogeneous one-dimensional habitat. The model took selection and migration into account and resulted in a recursion of the form

$$u_{n+1}=Q[u_n], \tag{1.1}$$

where $u_n(x)$ is the gene fraction of the population at location x in the nth generation. The operator

$$Q[u](x)=\int K(x-y)g(y,u(y))\,dy \tag{1.2}$$

is defined on the set of functions $\mathscr{C}=\{u: 0\leqq u\leqq 1,\ u \text{ piecewise continuous}\}$.

In the model, the selection process is described by a function $g:\mathbb{R}\times[0,1]\to[0,1]$, where

$$g(x,u)=\frac{s(x)u^2+u}{1+s(x)u^2+\sigma(x)(1-u)^2}. \tag{1.3}$$

Migration on the other hand is described by a probability density function K.

The formula (1.3) was arrived at under several severe restrictions, among which is the fact that fitnesses of the three genotypes AA, Aa and aa present in the population have to be in the ratio $1+s:1:1+\sigma$. In actual situations, the difference between these fitnesses is usually small.

Equation (1.1) has so far been studied only when $s\geqq\sigma$ are constants (g independent of x). The case $s>0>\sigma$ and $s\geqq\sigma>0$ are considered in the papers [10], [11] and [12], [13] respectively. The case $0>s\geqq\sigma$ is essentially the same as that of $s>0>\sigma$. It has also been mentioned in these papers that our model came as an improvement of a similar model proposed by R. A. Fisher in 1937 [6].

Fisher came up with the nonlinear diffusion equation $u_t=u_{xx}+f(u)$. This equation has received a lot of attention lately (see references in [13]). Our results in [10] through [13] agreed to a remarkable extent with those obtained for Fisher's equation. Not surprisingly, the results in this paper are in line with those in [4] and [18]. Judging from what is known, it is clear that the qualitative picture of the solutions is independent of the details of the modelling and therefore has much biological interest.

*Received by the editors January 17, 1984.
†Department of Mathematical Sciences, Worcester Polytechnic Institute, Worcester, Massachusetts 01609.

The purpose of this paper is to study (1.1) without assuming that s and σ are constants. We assume however that individuals of genotype AA are more fit to survive in the far left region of the habitat while the same is true for genotype aa in the far right. In terms of s and σ, we assume

(1.4) There exists $N>0$ such that $s(x)\geqq\sigma(x)$, $s(x)>0$ for $x\leqq -N$ and $s(x)\leqq\sigma(x)$, $\sigma(x)>0$ for $x\geqq N$.

We also assume that none of the homozygotes is lethal. That is to say,

$$1+s(x)>0, \qquad 1+\sigma(x)>0 \quad \text{in } \mathbb{R}. \tag{1.5}$$

This implies that $g(x,0)\equiv 0$, $g(x,1)\equiv 1$ and $\tilde{g}(x,0)\equiv 0$, $\tilde{g}(x,1)\equiv 1$. Here

$$\tilde{g}(x,u)\equiv 1-g(x,1-u)=\frac{\sigma(x)u^2+u}{1+\sigma(x)u^2+s(x)(1-u)^2}. \tag{1.6}$$

From (1.3)

$$g_x(x,u)=\frac{\left[s'u+(\sigma s'-s\sigma')u(1-u)-\sigma'(1-u)\right]u(1-u)}{\left[1+su^2+\sigma(1-u)^2\right]^2}.$$

According to (1.5), the denominator is always positive. If $\sigma'(x)>0$, $s'(x)<0$ and $0<u<1$, then $s'u[1+\sigma(1-u)]<0<\sigma'(1-u)[1+su]$ so that $g_x(x,u)<0$. Therefore we assume, in addition to (1.4) and (1.5),

$$\sigma'(x)\geqq 0, \qquad s'(x)\leqq 0 \quad \text{for } |x|\geqq N. \tag{1.7}$$

This will imply that $g_x(x,u)\leqq 0$ for $|x|\geqq N$.

Condition (1.7) is only enough to guarantee the existence of clines but not the uniqueness or stability. For these we need the more restrictive assumption

(1.8) $\sigma'(x)\geqq 0$, $s'(s)\leqq 0$ in $\mathbb{R}$ and there exists an interval $\mathscr{J}$ where $\sigma'(x)>0$, $s'(x)<0$.

In terms of g, (1.8) implies that, $g_x(x,u)\leqq 0$ and $g_x(x,u)<0$ in $\mathscr{J}\times(0,1)$.

Again from (1.3)

$$g_u(x,u)=\frac{-(s+2s\sigma+\sigma)u^2+(2s+2s\sigma)u+1+\sigma}{\left[1+su^2+\sigma(1-u)^2\right]^2}.$$

Let $N(u)=-(s+2s\sigma+\sigma)u^2+(2s+2s\sigma)u+1+\sigma$. Then $N(0)=1+\sigma>0$, $N(1)=1+s>0$ so that $N(u)>0$ in $[0,1]$ if $s+2s\sigma+\sigma\geqq 0$. When $s+2s\sigma+\sigma<0$, the minimum of N occurs at $u^*=s(1+\sigma)/(s+2s\sigma+\sigma)$. In order for $0\leqq u^*\leqq 1$, we must have $s\leqq 0$, $\sigma\leqq 0$. But then

$$N(u^*)=\frac{(1+\sigma)(1+s)[s+s\sigma+\sigma]}{s+2s\sigma+\sigma}>0,$$

since $s+s\sigma+\sigma\leqq s+2s\sigma+\sigma<0$. Thus $g_u(x,u)>0$ in $\mathbb{R}\times[0,1]$. Note that $g_u(x,0)=1/(1+\sigma)$ and $g_u(x,1)=1/(1+s)$.

To continue, consider

$$g(-N,u)=\frac{s(-N)u^2+u}{1+s(-N)u^2+\sigma(-N)(1-u)^2}.$$

From (1.4), $s(-N)$ is positive so that the numerator and denominator are both nonnegative. Since $s(-N)\geqq\sigma(-N)$, (1.7) implies that

$$(1.9)\quad g(x,u)\geqq g(-N,u)\geqq g_1(u)\equiv\frac{s(-N)u^2+u}{1+s(-N)u^2+s(-N)(1-u)^2}\quad\text{for } x\leqq -N.$$

Similarly, $\sigma(N)\geqq s(N)$ and (1.7) imply that

$$(1.10)\quad g(x,u)\leqq g(N,u)\leqq g_0(u)\equiv\frac{\sigma(N)u^2+u}{1+\sigma(N)u^2+\sigma(N)(1-u)^2}\quad\text{for } x\geqq N.$$

This last inequality is easy to verify if we look at $\tilde{g}(N,u)$.

We now summarize the hypotheses on K and g to be assumed throughout the entire paper except for condition (viii*) of (1.12). The hypotheses on K are identical to those assumed in [12]. We shall not assume g has the form (1.3) but only that it satisfies all the conditions listed in (1.12). Our discussion earlier made it clear what to assume of s and σ in order that (1.12) holds when g has the form (1.3).

(i) $K(x)\geqq 0$. If $B_1=\inf\{x:K(x)>0\}$, $B_2=\sup\{x:K(x)>0\}$, then $K(x)>0$ in (B_1,B_2). We allow $B_1=-\infty$, $B_2=+\infty$ so that K need not have compact support.

(ii) $K(x)$ is continuous in $\mathbb{R}$, except possibly at B_1 and B_2 where $\lim_{x\downarrow B_1}K(x)=p_1$, $\lim_{x\uparrow B_2}K(x)=p_2$.

(1.11) Also K may be written in the form

$$K(x)=K_a(x)-p_1\chi_{(-\infty,B_1]}-p_2\chi_{[B_2,\infty)},$$

where K_a is absolutely continuous and χ_S is the indicator function of the set S.

(iii) $\int K(x)\,dx=1$.

(iv) $\int e^{\mu x}K(x)\,dx$ is finite for all real μ.

(v) $\int_x^\infty K(y)\,dy\leqq \text{const}\,K(x)$ for large x, $\int_{-\infty}^x K(y)\,dy\leqq\text{const}\,K(x)$ for small x.

(vi) $g(x,u):\mathbb{R}\times[0,1]\to[0,1]$ has continuous derivative. g_x, g_u, g_{uu} are uniformly bounded.

(vii) $g(x,0)\equiv 0$, $g(x,1)\equiv 1$.

(viii) There exists $N>0$ such that $g_x(x,u)\leqq 0$ for $|x|\geqq N$; or

(viii*) $g_x(x,u)\leqq 0$ in $\mathbb{R}\times[0,1]$ and $g_x(x,u)<0$ in $\mathscr{J}\times(0,1)$ for some interval $\mathscr{J}$.

(ix) $g_u(x,u)\geqq 0$ in $\mathbb{R}\times[0,1]$ and $g_u\not\equiv 0$ in any rectangle.

(1.12) (x) $g_u(x,0)\in(0,1)$ uniformly for $x\geqq N$, $g_u(x,1)\in(0,1)$ uniformly for $x\leqq -N$.

(xi) There exist two functions g_+, g_- satisfying all the conditions in (1.13) such that $g(-N,u)\geqq g_-(u)$ and $g(N,u)\leqq g_+(u)$. Furthermore, $c_+^*(g_-)>0$ and $c_-^*(g_+)>0$.

At this point we must digress to explain the meaning of condition (xi). Consider first a function $g:[0,1]\to[0,1]$ satisfying the conditions:

(1.13)
(i) $g\in C^1[0,1]$.
(ii) $g(0)=0$, $g(1)=1$.
(iii) There exists a constant $\alpha\in(0,1)$ such that $g(u)<u$ in $(0,\alpha)$ and $g(u)>u$ in $(\alpha,1)$.
(iv) $g'(u)\geqq 0$ in $[0,1]$. If $\sigma_1=\inf\{u:g(u)>0\}$, $\sigma_2=\sup\{u:g(u)<1\}$, then $g'(u)>0$ in (σ_1,σ_2).
(v) $g'(0)<1$, $g'(1)<1$.
(vi) $g(u)\geqq g'(\alpha)(u-\alpha)+\alpha$ in $[0,\alpha]$ and $g(u)\leqq g'(\alpha)(u-\alpha)+\alpha$ in $[\alpha,1]$.
(vii) $g'(0)u\leqq g(u)\leqq g'(1)(u-1)+1$ in $[0,1]$.

Remark 1.1. Condition (vi) implies that $\max_{[0,1]}g(u)/u<g'(\alpha)$ and $\max_{[0,1]}((1-g(1-u))/u)<g'(\alpha)$. All the results in [12] are valid under conditions (1.11) and (1.13).

Let $\bar{Q}:\mathscr{C}\to\mathscr{C}$ be the nonlinear convolution operator $\bar{Q}[u]=K*g(u)$. Then associated with $\bar{Q}$ is a real number $c_+^*(g)$ such that the following holds.

THEOREM 1.1 (existence of travelling waves). *There exists a nonincreasing function* w, $w(-\infty)=1$, $w(\infty)=0$ *and* $w(x)=\bar{Q}[w](x+c_+^*(g))$.

THEOREM 1.2 (uniqueness). *Suppose* $u\in\mathscr{C}$ *satisfies* $u(-\infty)>\alpha$, $u(\infty)<\alpha$ *and* $u(x)=\bar{Q}[u](x+c)$. *Then* $c=c_+^*(g)$ *and* $u(x)=w(x-\tau)$ *for some constant* τ.

The above two theorems are [12, Thm. 5] and [2, Thm. 1.2] respectively. The function w is called a travelling wave solution of $\bar{Q}$ facing right. They are unique up to translation.

The number $c_+^*(g)$ is called the wave speed of $\bar{Q}$ in the positive direction [13], [21]. It should be pointed out that $c_+^*(g)$ is the asymptotic speed of propagation for certain class of initial data. For example, let $u_0\in\mathscr{C}$ be decreasing, $u_0(-\infty)>\alpha$, $u_0(\infty)<\alpha$ and $u_{n+1}=\bar{Q}[u_n]$ for all n. Then $\lim_{n\to\infty}(u_n^{-1}(\gamma)/n)=c_+^*(g)$ for every $0<\gamma<1$.

There are of course nondecreasing travelling waves facing left with wave speed $c_-^*(g)$ in the negative direction. The meaning of condition (xi) should now be clear.

Remark 1.2. If $u_n(x)=0$ for $x\geqq 0$, then $u_{n+1}(x)=0$ for $x\geqq B_2$. Thus the speed of propagation to the right, namely $c_+^*(g)$, cannot exceed B_2. We can show that $B_1<c_+^*(g)<B_2$ and $B_1<-c_-^*(g)<B_2$ so that condition (xi) implies $B_1<0<B_2$.

Remark 1.3. The wave speed depends monotonically on $\bar{Q}$. Given $\bar{Q}_1$ and $\bar{Q}_2$ with $\bar{Q}_1[u]\geqq\bar{Q}_2[u]$ for all $u\in\mathscr{C}$, then $c_+^*(\bar{Q}_1)\geqq c_+^*(\bar{Q}_2)$. For example, if $g_1(u)\geqq g_2(u)$ in $[0,1]$ and $\bar{Q}_1[u]=K*g_1(u)$, $\bar{Q}_2[u]=K*g_2(u)$, then $c_+^*(g_1)\geqq c_+^*(g_2)$. In fact, more is true. If $g_1(u)>g_2(u)$ in $(0,1)$ and $g_1'(0)>g_2'(0)$, $g_1'(1)<g_2'(1)$, then $c_+^*(g_1)>c_+^*(g_2)$ [2].

Remark 1.4. Let g_1, g_0 be defined as in (1.9) and (1.10). Then all the conditions in (1.13) are satisfied. It is easy to check that $\alpha=1/2$, $\sigma_1=0$, $\sigma_2=1$, $g_1'(0)=g_1'(1)=(1+s(-N))^{-1}$, $g_0'(0)=g_0'(1)=(1+\sigma(N))^{-1}$ and $g_i'(u)\leqq g_i'(1/2)$ in $[0,1]$. This last inequality obviously implies condition (vi) of (1.13). It may be proved by observing that $g_i'(u)$ is a rational function in u. The numerator has a maximum at $u=1/2$ and the denominator has a minimum also at $u=1/2$. The left-hand inequality in (vii) of (1.13) is straightforward. The right-hand inequality is equivalent to showing that $\tilde{g}_i(u)\equiv 1-g_i(1-u)\geqq g_i'(1)u$. But then $\tilde{g}_i=g_i$ and $g_i'(0)=g_i'(1)$ from (1.6), and so it is the same as the left-hand inequality.

Remark 1.5. If g is given by (1.3), we cannot take $g_-=g_1$ or $g_+=g_0$ in assumption (xi) of (1.12). In fact when K is even, $c_+^*(g_1)=c_-^*(g_0)=0$ [2]. However, if $s(-N)>\sigma(-N)$ in (1.4), we let $g_-(u)=(s(-N)u^2+u)/(1+s(-N)u^2+\sigma_1(1-u)^2)$ for some

$\max\{\sigma(-N),0\}<\sigma_1<s(-N)$. Then $g(-N,u)\geqq g_-(u)$ and if σ_1, is close enough to $s(-N)$, (1.13) will be satisfied by $c_+^*(g_-)>0$ (Remark 1.3). A similar arrangement can be made for g_0 when $s(N)<\sigma(N)$.

Finally a few words about references. As mentioned earlier, results in this paper parallel those obtained in [4] which considered the differential equation $u_t=u_{xx}+f(x,u)$. Almost identical results were obtained in [18] for the equation $u_t=u_{xx}+mu_x+f(x,u)$. The term mu_x came from assuming nonsymmetric migration. In [5], one of the homozygotes was assumed favored in the entire habitat, and the differential equation was allowed to have variable coefficients in some cases. This could happen in our model also if we do not assume the total population density $\mu(x)$ is a constant. Then K in (1.2) is replaced by

$$K(x,y)=\frac{K(x-y)\mu(y)}{\int K(x-y)\mu(y)\,dy},$$

see [21]. It is not clear if any of the techniques developed so far are applicable to this case.

The paper by Felsenstein [3] contains 152 references on the subject of variable selection and migration. Some of the fairly standard ones are [7], [14]–[17], [20]. We must also mention the work of Conley [1], who proved the existence of clines for the above differential equation with $f(x,u)=s(x)u(1-u)$, $s(\pm\infty)\neq 0$, using a topological argument. A radially symmetric problem with $x\in\mathbb{R}^2$ is also considered. The paper by Sawyer [19] contains more complete and recent information.

2. Statement of results.

THEOREM 2.1. *There exists a function* $\phi\in\mathscr{C}$ *such that* $\phi(-\infty)=1$, $\phi(\infty)=0$ *and* $\phi=Q[\phi]$.

Remark 2.1. If ϕ is nonincreasing, it is often referred to as a cline.

The rest of the results assume the stronger condition (viii*) of (1.12).

THEOREM 2.2. *There is at most one solution to the problem* $\phi\in\mathscr{C}$, $\phi(-\infty)=1$, $\phi(\infty)=0$ *and* $\phi=Q[\phi]$. *Furthermore, such a solution is decreasing in* $\mathbb{R}$.

In the next two theorems, u_n is defined recursively by (1.1) for a given u_0. ϕ is the unique monotone cline from Theorems 2.1 and 2.2.

From conditions (viii*) and (xi) of (1.12), we have $g(x,u)\geqq g_-(u)$ for $x\leqq -N$ and $g(x,u)\leqq g_+(u)$ for $x\geqq N$. Since $g_-(u)>u$ near 1 and $g_+(u)<u$ near 0, we can define the functions $a_1(x)=\inf\{u:g(x,u)>u\}$ for $x\leqq -N$ and $a_0(x)=\sup\{u:g(x,u)<u\}$ for $x\geqq N$. Also, $a_1(x)$ is nondecreasing in x and a_0 is nonincreasing in x. We let $a^-=\lim_{x\to\infty}a_0(x)$ and $a^+=\lim_{x\to-\infty}a_1(x)$. Clearly $a^+<1$ and $a^->0$.

THEOREM 2.3. *Suppose* u_0 *satisfies the condition* (i) $\phi(x-h_1)\leqq u_0(x)\leqq\phi(x-h_2)$ *for some* $h_1<0<h_2$ *or* (ii) $\liminf_{x\to-\infty}u_0(x)>a^+$, $\limsup_{x\to\infty}u_0(x)<a^-$, *then* $\lim_{n\to\infty}\|u_n-\phi\|_\infty=0$.

THEOREM 2.4. *There exist positive constants* δ, μ *and* C *such that if* $\|u_0-\phi\|_\infty\leqq\delta$, *then* $\|u_n-\phi\|_\infty\leqq Ce^{-\mu n}$ *for all* n. *Consequently, the uniform convergence in Theorem* 2.3 *may be replaced by exponential convergence.*

3. Proof of Theorem 2.1. We begin by proving left continuity of $c_+^*(g)$ with respect to g. It is a consequence of Theorem 1.2.

LEMMA 3.1. *Suppose* g_δ, $0<\delta<\delta_0$ *is a family of functions each of which satisfies the conditions in* (1.13) *with some* $\alpha_\delta\in(0,1)$. *Suppose further that* g_δ *increases uniformly to* g *as* $\delta\downarrow 0$. *Then* $\lim_{\delta\downarrow 0}c_+^*(g_\delta)=c_+^*(g)$.

Proof. By Remark 1.3, $c_\delta \equiv c_+^*(g_\delta)$ increases as δ decreases. Let $\lim_{\delta \downarrow 0} c_\delta = c_0 \leqq c_+^*(g)$. From Theorem 1.1, for each $\delta > 0$, there exists nonincreasing travelling waves w_δ such that $w_\delta(-\infty)=1$, $w_\delta(\infty)=0$ and

$$w_\delta(x)=\int K(x+c_\delta-y)g_\delta(w_\delta(y))\,dy.$$

Since w_δ is determined only up to translation, we may choose w_δ such that $w_\delta(0)=\gamma$ for some fixed $\gamma \in (\alpha,1)$. From (ii) of (1.11), $\|w_\delta'\|_\infty \leqq \|K_a'\|_1 + p_1 + p_2$. Arzela–Ascoli theorem implies that a subsequence, also denoted by w_δ, converges uniformly on compact sets to a nonincreasing function w^*. Furthermore, $w^*(0)=\gamma$ and $w^*(x)=\int K(x+c_0-y)g(w^*(y))\,dy$. Therefore $w^*(-\infty)=1$ and $w^*(\infty)=\alpha$ or 0. $w^*(\infty)$ cannot be α because $g(u)>u$ in $(\alpha,1)$ and such a solution connecting 1 and α exists if and only if $c_0 \geqq \rho_+^* > c_+^*(g)$. See [12, Prop. 3 and Lemma 2.2] for this fact and the definition of ρ_+^*. Thus $w^*(\infty)=0$ and by Theorem 1.2, $c_0 = c_+^*(g)$. Q.E.D.

LEMMA 3.2. *There exist two nonincreasing functions $\underline{u}$, $\bar{u}$ in $\mathscr{C}$ with the properties $\underline{u} \leqq \bar{u}$, $\underline{u} \leqq Q[\underline{u}]$ and $Q[\bar{u}] \leqq \bar{u}$.*

Proof. We first construct $\underline{u}$. From condition (xi) of (1.12), there exists g_- satisfying (1.13) with $c_+^*(g_-)>0$. Let $\delta_0>0$ be sufficiently small and for each $0<\delta<\delta_0$, construct $g_\delta \in C^1[0,1]$ such that $g_\delta = g_-$ on $[\delta,1]$, $g_\delta = 0$ on $[0,\delta/2]$ and on the interval $(\delta/2,\delta)$, $g_\delta'>0$ and g_δ increases uniformly to g_- as $\delta \downarrow 0$. It is clear that (1.13) is satisfied for each g_δ.

From Lemma 3.1, for sufficiently small $\delta>0$, $c_\delta \equiv c_+^*(g_\delta)>0$. Fix such a δ and let w_δ be the nonincreasing travelling wave of the operator $K * g_\delta(u)$, translated so that $w_\delta(-N)=\delta/2$. We have

$$\begin{aligned} w_\delta(x) &= \int K(x+c_\delta-y)g_\delta(w_\delta(y))\,dy = \int K(x-y)g_\delta(w_\delta(y+c_\delta))\,dy \\ &\leqq \int K(x-y)g_\delta(w_\delta(y))\,dy = \int_{-\infty}^{-N} K(x-y)g_\delta(w_\delta(y))\,dy \\ &\leqq \int_{-\infty}^{-N} K(x-y)g_-(w_\delta(y))\,dy. \end{aligned}$$

Let

$$\underline{u}(x)=\begin{cases} w_\delta(x) & \text{if } x \leqq -N, \\ 0 & \text{if } x > -N. \end{cases}$$

Then from (1.2) and (viii) of (1.12), we have

$$\begin{aligned} Q[\underline{u}](x) &= \int_{-\infty}^{-N} K(x-y)g(y,\underline{u}(y))\,dy \geqq \int_{-\infty}^{-N} K(x-y)g(-N,\underline{u}(y))\,dy \\ &\geqq \int_{-\infty}^{-N} K(x-y)g_-(\underline{u}(y))\,dy \geqq w_\delta(x). \end{aligned}$$

Therefore, $Q[\underline{u}](x) \geqq \underline{u}(x)$ for $x \leqq -N$. Since $Q[\underline{u}] \geqq 0$, the inequality holds for all x.

To construct $\bar{u}$, let $K_1(x)=K(-x)$, $\tilde{g}_+(u)=1-g_+(1-u)$ and $\tilde{Q}[u]=K_1 * \tilde{g}_+(u)$. The relation between this (dual) operator $\tilde{Q}$ and $\overline{Q}[u]=K * g_+(u)$ is given in [12, §2]. It is shown there that $\tilde{g}_+$ satisfies the set of hypotheses (1.13) and that the wave speed of $\tilde{Q}$ in the positive direction, hereby denoted by $\tilde{c}_+^*(\tilde{g}_+)$, is equal to the wave speed of $\overline{Q}$ in the negative direction, $c_-^*(g_+)$. This fact is a consequence of the symmetry between 0 and 1 in the graph of g_+.

As before, we construct $\tilde{g}_\delta$ increasing uniformly to $\tilde{g}_+$ as $\delta \downarrow 0$, $\tilde{g}_\delta(u)=0$ in $[0,\delta/2]$, $\tilde{g}_\delta=\tilde{g}_+$ in $[\delta,1]$ and each $\tilde{g}_\delta$ satisfies (1.13). Since $c^*_-(g_+)>0$, Lemma 3.1 and above implies that for sufficiently small $\delta>0$, $\tilde{c}_\delta \equiv \tilde{c}^*_+(\tilde{g}_\delta)$ is positive.

Now let $\tilde{w}_\delta$ be the nonincreasing travelling wave of the operator $K_1 * \tilde{g}_\delta(u)$ translated so that $\tilde{w}_\delta(-N)=\delta/2$. As before,

$$\begin{aligned}\tilde{w}_\delta(x) &= \int K_1(x-y)\tilde{g}_\delta(\tilde{w}_\delta(y+\tilde{c}_\delta))\,dy \\ &\leqq \int_{-\infty}^{-N} K(-x+y)\tilde{g}_\delta(\tilde{w}_\delta(y))\,dy \\ &= \int_N^\infty K(-x-y)\tilde{g}_\delta(\tilde{w}_\delta(-y))\,dy.\end{aligned}$$

Define $v(x)=1-\tilde{w}_\delta(-x)$. We have

$$\begin{aligned}v(x) &\geqq 1-\int_N^\infty K(x-y)\tilde{g}_\delta(\tilde{w}_\delta(-y))\,dy \\ &= 1-\int_N^\infty K(x-y)\tilde{g}_\delta(1-v(y))\,dy.\end{aligned}$$

Now let

$$\bar{u}(x)=\begin{cases}1 & \text{if } x\leqq N,\\ v(x) & \text{if } x\geqq N,\end{cases}$$

$\tilde{g}(x,u)=1-g(x,1-u)$. Then

$$\begin{aligned}Q[\bar{u}](x) &= 1-\int K(x-y)\tilde{g}(y,1-\bar{u}(y))\,dy \\ &= 1-\int_N^\infty K(x-y)\tilde{g}(y,1-\bar{u}(y))\,dy.\end{aligned}$$

From condition (xi) of (1.12), $\tilde{g}(x,u)\geqq \tilde{g}_+(u)$ for $x\geqq N$ so that

$$\begin{aligned}Q[\bar{u}](x) &\leqq 1-\int_N^\infty K(x-y)\tilde{g}_+(1-\bar{u}(y))\,dy \\ &\leqq 1-\int_N^\infty K(x-y)\tilde{g}_\delta(1-\bar{u}(y))\,dy \leqq v(x).\end{aligned}$$

Since $\bar{u}(x)=v(x)$ for $x\geqq N$ and $Q[\bar{u}]\leqq 1$, we have $Q[\bar{u}]\leqq \bar{u}$. It is also clear from the definitions of $\underline{u}$ and $\bar{u}$ that they are nonincreasing and $\underline{u}\leqq\bar{u}$. This completes the proof of Lemma 3.2.

To prove Theorem 2.1, we first observe that Q is order-preserving in the sense that $u\leqq v$ implies $Q[u]\leqq Q[v]$.

Let $\underline{u}_0=\underline{u}$, $\bar{u}_0=\bar{u}$ and define $\underline{u}_n$, $\bar{u}_n$ recursively as in (1.1). An inductive argument shows that $\underline{u}_0\leqq\underline{u}_n\leqq\underline{u}_{n+1}\leqq\bar{u}_{n+1}\leqq\bar{u}_n\leqq\bar{u}_0$ for all n. Therefore $\underline{u}_n$, $\bar{u}_n$ converge, as $n\to\infty$, to ϕ_1, ϕ_2 respectively. Since $\underline{u}\leqq\phi_1\leqq\phi_2\leqq\bar{u}$, we have $\phi_i(-\infty)=1$, $\phi_i(\infty)=0$ and $\phi_i=Q[\phi_i]$. The proof of Theorem 2.1 is complete if we take $\phi=\phi_1$ or ϕ_2.

Remark 3.1. It does not follow from the above construction that ϕ_i is nonincreasing in $\mathbb{R}$, even though $\underline{u}$, $\bar{u}$ are. This is true if we assume $g_x\leqq 0$. In this case there exists a cline.

Remark 3.2. The solution ϕ we have constructed in Theorem 2.1 lies between $\underline{u}$ and $\bar{u}$. Since $w_\delta(x)$, $\tilde{w}_\delta(x)$ converge to 1 exponentially as $x \to -\infty$ [12, Prop. 5], there exist $\mu > 0$, $C > 0$ such that $|\phi(x)| \leqq Ce^{-\mu x}$ as $x \to \infty$ and $|1-\phi(x)| \leqq Ce^{\mu x}$ as $x \to -\infty$.

Without further assumptions other than (viii) of (1.12), the solution ϕ is not unique. For example, let $h(x,u)$, defined on $\mathbb{R} \times [0,1]$, be sufficiently smooth and satisfy the conditions (i) $h(x,u)=0$ on $(-\infty, N] \times [\frac{1}{2}, 1]$ and $[-N, \infty) \times [0, \frac{1}{2}]$, (ii) $h(x,0)=h(x,1)=0$, (iii) $h_x(x,u) \leqq 0$ for $|x| \geqq N$ and (iv) $-\frac{1}{2} \leqq h(x,u) \leqq \frac{1}{2}$ in $\mathbb{R} \times [0,1]$. It is clear that such an h exists and if $\gamma \in (0,1)$, γh also satisfies conditions (i) to (iv).

Let $K(x)=K(-x)$, $g(x,u)=g_0(u)+\gamma h(x,u)$, where g_0 is given by (1.10). Choose γ so small that $g_0'(u)+\gamma h_u(x,u) > 0$ in $\mathbb{R} \times [0,1]$. This is possible since $\min_{[0,1]} g_0'(u) > 0$. For small γ, it is straightforward to verify that conditions (vi) to (x) of (1.12) are satisfied for g. However, according to Theorem 1.1 and Remark 1.5, there exists a nonincreasing function w, $w(-\infty)=1$, $w(0)=1/2$ and $w(\infty)=0$ such that $w(x)=\int K(x-y) g_0(w(y))\,dy$. From (i) above, it is easy to see that $w(x+\tau)=\int K(x-y) g(y, w(y+\tau))\,dy$ for $|\tau| \leqq N$. Therefore we have nonuniqueness.

4. Proof of Theorem 2.2. For the rest of the paper we assume condition (viii*) of (1.12). Proposition 4.2 is the basis for much of the results that follow. The following lemma is the heart of its proof.

LEMMA 4.1. *Let ϕ be nonincreasing, $\phi(-\infty)=1$, $\phi(\infty)=0$ and $\phi = Q[\phi]$. There exist two decreasing sequences $\{z_n\}$, $\{q_n\}$ such that if $v_n(x)=\phi(x-z_n)-q_n$, then $v_{n+1} \leqq Q[v_n]$ for all n.*

Remark 4.1. We shall see from the proof that there are no restrictions on z_0, q_0 except that $z_0 \leqq 0$ and $q_0 > 0$ be sufficiently small.

Proof. We begin by showing $\phi'(x) < 0$ in $\mathbb{R}$. From $g_x \leqq 0$ and our hypotheses, $\phi'(x) \leqq \int K(x-y) g_u(y, \phi(y)) \phi'(y)\,dy \leqq 0$. Let $\phi'(x_0)=0$. Then $g_u(y, \phi(y))\phi'(y)=0$ on the interval $[x_0-B_2, x_0-B_1]$, which, according to Remark 1.2, contains x_0. Since g_u does not vanish on any rectangle, $\phi'(x)=0$ on an open interval containing x_0. This means that the set S when $\phi'=0$ is open. From the continuity of ϕ', S is closed. Obviously ϕ is not a constant, S is empty and so $\phi'(x) < 0$ in $\mathbb{R}$.

Next we show that there exist constants q_0, δ, θ_1 all in the interval $(0,1)$ such that

$$g(x,u-q)-g(x,u) \geqq -\theta_1 q \quad \text{for } 0 \leqq q \leqq q_0, \tag{4.1}$$
$$u \in [1-\delta, 1] \text{ and } x \leqq -N \quad \text{or} \quad u \in [0, \delta] \text{ and } x \geqq N.$$

To begin, consider the function

$$\psi(x,u,q) = \begin{cases} \dfrac{g(x,u)-g(x,u-q)}{q} & \text{if } q > 0, \\ g_u(x,u) & \text{if } q = 0 \end{cases}$$

in the set $\mathscr{D} = \mathbb{R} \times [1-\delta, 1] \times [0, q_0]$, where we shall define $g(x,u)=0$ if $u < 0$. It is clear that ψ is uniformly continuous in $\mathscr{D}$. Also

$$\psi(x,1,q) = \frac{1-g(x,1-q)}{q} = g_u(x,\theta) \quad \text{where } 1-q \leqq \theta \leqq 1.$$

From (x) of (1.12) and the fact that g_u is uniformly continuous, there exists $\theta_1 \in (0,1)$ such that $\psi(x,1,q) < \theta_1 < 1$ for q sufficiently small and $x \leqq -N$. Since ψ is uniformly continuous in $\mathscr{D}$, $\psi(x,u,q) < \theta_1$ for u near 1. Therefore (4.1) holds when $x \leqq -N$ and $1-\delta \leqq u \leqq 1$.

Next consider $g(x,u-q)-g(x,u)$ for $0\leqq u\leqq\delta$ and $x\geqq N$. If $u-q\geqq 0$, then $g(x,u-q)-g(x,u)=-g_u(x,\theta)q$, where $u-q\leqq\theta\leqq u$. From (x) of (1.12) and uniformly continuity of g_u, we may assume (by increasing $\theta_1<1$ if necessary) that $g_u(x,\theta)<\theta_1<1$ for $x\geqq N$ and θ sufficiently small. Therefore (4.1) holds when $x\geqq N$ and u sufficiently small.

On the other hand, if $u-q<0$, then $g(x,u-q)-g(x,u)=-g(x,u)$. From the mean value theorem and the fact that g_{uu} is bounded, $\lim_{u\downarrow 0}(g(x,u)/u)=g_u(x,0)$ uniformly in $\mathbb{R}$. Therefore for $x\geqq N$ and u small $g(x,u)/u<\theta_1$ so that $g(x,u-q)-g(x,u)\geqq-\theta_1 u\geqq-\theta_1 q$. Altogether (4.1) is valid. It should be pointed out that (4.1) continues to hold with the same θ_1 and N if we decrease q_0 and δ.

To continue, let $M=\sup_{\mathbb{R}\times[0,1]}g_u(x,u)\geqq 1$ and choose $\varepsilon>0$, $\eta>0$ such that $\theta=\varepsilon M+\theta_1<1$,

$$\int_{\eta}^{\infty}K(x)\,dx\leqq\varepsilon,\qquad\int_{-\infty}^{-\eta}K(x)\,dx\leqq\varepsilon. \tag{4.2}$$

Define μ and q_n by $\theta=e^{-\mu}$ and $q_n=q_0e^{-\mu n}=q_0\theta^n$ for all n. Since $\phi'<0$ in $\mathbb{R}$, $\phi(-\infty)=1$, $\phi(\infty)=0$, we define $E_\gamma=\phi^{-1}(\gamma)$ for every $0<\gamma<1$.

Let $\Gamma=[E_{1-\delta}-2\eta,\ E_\delta+2\eta]$. We assume $\delta>0$ is sufficiently small so that $E_\delta\geqq N$, $E_{1-\delta}\leqq -N$. There exists $\theta_2>0$ such that

$$\phi(\xi_1)-\phi(\xi_2)\leqq-\theta_2(\xi_1-\xi_2)\quad\text{if }\xi_1>\xi_2\text{ are in }\Gamma. \tag{4.3}$$

Finally, let $z_0\leqq 0$ be arbitrary and define z_n recursively by

$$z_{n+1}=\frac{(\theta-M)q_0e^{-\mu n}}{\theta_2}+z_n. \tag{4.4}$$

Clearly z_n's are nonpositive and decreasing and converge to the limit

$$x_1=\frac{(\theta-M)q_0}{\theta_2}\left(\frac{1}{1-e^{-\mu}}\right)+z_0.$$

We may assume that q_0 is sufficiently small so that $(M-\theta)q_0\theta_2^{-1}\leqq\eta$.

Having defined all the constants, we proceed to prove the inequality $v_{n+1}\leqq Q[v_n]$, where $v_n(x)=\phi(x-z_n)-q_n$. This is equivalent to showing

$$\begin{aligned}&\phi(x-z_{n+1})-\phi(x-z_n)\\&\quad-\int K(x-y)\left[g(y,\phi(y-z_n)-q_n)-g(y-z_n,\phi(y-z_n))\right]dy\leqq q_{n+1}.\end{aligned} \tag{4.5}$$

Let

$$\Gamma_n=[E_{1-\delta}+z_n,\ E_\delta+z_n],\qquad\Gamma_n'=[E_{1-\delta}+z_n-\eta,\ E_\delta+z_n+\eta]$$

and

$$h_n(x)=g(x,\phi(x-z_n)-q_n)-g(x-z_n,\phi(x-z_n)).$$

Then

$$\begin{aligned}\int K(x-y)h_n(y)\,dy&=\int_{\Gamma_n}K(x-y)h_n(y)\,dy+\int_{y\geqq E_\delta+z_n}K(x-y)h_n(y)\,dy\\&\quad+\int_{y\leqq E_{1-\delta}+z_n}K(x-y)h_n(y)\,dy,\\&\equiv I_1+I_2+I_3.\end{aligned}$$

Consider first the case $x \notin \Gamma_n'$. Since $z_n \leqq 0$,

$$I_1 \geqq \int_{\Gamma_n} K(x-y)\left[g(y,\phi(y-z_n)-q_n)-g(y,\phi(y-z_n))\right] dy$$

$$= \int_{\Gamma_n} K(x-y)\, g_u(y,\theta)(-q_n)\, dy \geqq -Mq_n \int_{\Gamma_n} K(x-y)\, dy \geqq -Mq_n\varepsilon.$$

The last inequality follows from (4.2).

If $y \geqq E_\delta + z_n$, then $y - z_n \geqq E_\delta \geqq N$ and $\phi(y-z_n) \leqq \delta$. From (4.1),

$$h_n(y) \geqq g(y-z_n, \phi(y-z_n)-q_n) - g(y-z_n, \phi(y-z_n)) \geqq -\theta_1 q_n.$$

Therefore, $I_2 \geqq -\theta_1 q_n \int_{y \geqq E_\delta + z_n} K(x-y)\, dy$. Similarly, if $y \leqq E_{1-\delta} + z_n$, then $y - z_n \leqq E_{1-\delta} \leqq -N$ and $\phi(y-z_n) \geqq 1-\delta$. Again from (4.1), $h_n(y) \geqq g(y, \phi(y-z_n)-q_n) - g(y, \phi(y-z_n)) \geqq -\theta_1 q_n$ so that $I_3 \geqq -\theta_1 q_n \int_{y \leqq E_{1-\delta}+z_n} K(x-y)\, dy$.

Combining all three inequalities, we have, when $x \notin \Gamma_n'$,

$$\int K(x-y) h_n(y)\, dy \geqq -Mq_n\varepsilon - \theta_1 q_n = -\theta q_n = -q_{n+1}.$$

Since $z_{n+1} \leqq z_n$ and ϕ is nonincreasing, the difference between the first two terms in (4.5) is nonpositive. Therefore (4.5) is established when $x \notin \Gamma_n'$.

If $x \in \Gamma_n'$, then $E_{1-\delta} - \eta \leqq x - z_n \leqq E_\delta + \eta$ and

$$h_n(y) \geqq g(y, \phi(y-z_n)-q_n) - g(y, \phi(y-z_n)) = g_u(y,\theta)(-q_n) \geqq -Mq_n.$$

Therefore $\int K(x-y) h_n(y)\, dy \geqq -Mq_n$. From (4.4), $z_n - z_{n+1} \leqq (M-\theta) q_0 \theta_2^{-1} \leqq \eta$ and hence $x - z_n \leqq x - z_{n+1} \leqq E_\delta + 2\eta$. From (4.3) and (4.4), $\phi(x-z_{n+1}) - \phi(x-z_n) \leqq -\theta_2(z_n - z_{n+1}) = (\theta - M) q_0 e^{-\mu n} = (\theta - M) q_n$. Hence (4.5) is valid if $x \in \Gamma_n'$. This completes the proof of Lemma 4.1.

PROPOSITION 4.2. *Let u_0 satisfy the conditions*

$$\liminf_{x \to -\infty} u_0(x) > a^+ \quad \text{and} \quad \limsup_{x \to \infty} u_0(x) < a^-.$$

Let ϕ be a nonincreasing function satisfying $\phi(-\infty) = 1$, $\phi(\infty) = 0$ and $\phi = Q[\phi]$. Then there exist constants x_1, x_2, q_0', μ, the last two positive, such that

$$\phi(x - x_1) - q_0' e^{-\mu n} \leqq u_n(x) \leqq \phi(x - x_2) + q_0' e^{-\mu n} \quad \text{for all } n.$$

Proof. We only prove the left-hand inequality. The right-hand inequality is the same but requires a result like Lemma 4.1 with $v_{n+1} \geqq Q[v_n]$. We begin by showing that $u_n(-\infty)$ increases to 1 as $n \to \infty$.

Since $a_1(x)$ decreases to a^+ as $x \to -\infty$ (see §2), there exist $\varepsilon > 0$, $N_\varepsilon > N$ such that

$$a^+ \leqq a_1(x) \leqq a_1(-N_\varepsilon) < a^+ + \varepsilon \leqq u_0(x)$$

and

$$g(x, u_0(x)) \geqq g(-N_\varepsilon, u_0(x)) \geqq g(-N_\varepsilon, a^+ + \varepsilon) \quad \text{for } x \leqq -N_\varepsilon.$$

From Fatou's lemma,

$$\liminf_{x \to -\infty} u_1(x) \geqq \int K(y) \liminf_{x \to -\infty} g(x-y, u_0(x-y))\, dy \geqq g(-N_\varepsilon, a^+ + \varepsilon).$$

Now suppose $\liminf_{x \to -\infty} u_n(x) \geqq g^n(-N_\varepsilon, a^+ + \varepsilon)$. Then for any $\delta > 0$, $u_n(x) \geqq q^n(-N_\varepsilon, a^+ + \varepsilon) - \delta$ and $g(x, u_n(x)) \geqq g(-N_\varepsilon, u_n(x)) \geqq g(-N_\varepsilon, g^n(-N_\varepsilon, a^+ + \varepsilon) - \delta)$ for x

near $-\infty$. Hence $\liminf_{x\to-\infty} u_{n+1}(x) \geqq \int K(y) \liminf_{x\to-\infty} g(x-y, u_n(x-y))\,dy \geqq g(-N_\varepsilon, g^n(-N_\varepsilon, a^+ + \varepsilon) - \delta)$. Since $\delta > 0$ is arbitrary, we have

$$\liminf_{x\to-\infty} u_n(x) \geqq g^n(-N_\varepsilon, a^+ + \varepsilon) \quad \text{for all } n. \tag{4.6}$$

From the definition of $a_1(x)$, we have $g(-N_\varepsilon, u) > u$ for $a_1(-N_\varepsilon) < u < 1$. This implies that $g^n(-N_\varepsilon, a^+ + \varepsilon)$ increases to 1 as $n \to \infty$. Therefore, let $q_0 > 0$ be as defined in Lemma 4.1. There exist, by (4.6) positive integers k and N_0 such that $u_k(x) \geqq 1 - q_0$ for $x \leqq -N_0$. Hence $v_0(x) = \phi(x - z_0) - q_0 \leqq 1 - q_0 \leqq u_k(x)$ for $x \leqq -N_0$. From Remark 4.1, $z_0 \leqq 0$ is arbitrary and we now choose it sufficiently negative so that $\phi(-N_0 - z_0) - q_0 \leqq 0$. Therefore, $v_0(x) \leqq u_k(x)$ for all x.

Since $v_{n+1} \leqq Q[v_n]$ and Q is order-preserving, an inductive argument shows that $v_n \leqq u_{k+n}$ for all $n \geqq 0$. Explicitly,

$$\phi(x - z_n) - q_0 e^{-\mu n} \leqq u_{k+n}(x) \quad \text{for } n \geqq 0.$$

Since z_n is decreasing, we may replace z_n by its limit x_1 in the above inequality. Doing so and writing n for $k+n$, we have

$$\phi(x - x_1) - q_0' e^{-\mu n} \leqq u_n(x) \quad \text{for } n \geqq k, \text{ where } q_0' = q_0 e^{\mu k}.$$

The first k terms are then taken care of by increasing q_0' until $1 - q_0' e^{-\mu k} \leqq 0$. This completes the proof of Proposition 4.2.

Remark 4.2. The condition $\limsup_{x\to\infty} u_0(x) < a^-$ is needed to prove the right-hand inequality.

LEMMA 4.3. *Let ϕ satisfy $\phi = Q[\phi]$ and let $u_0(x) = \phi(x - h)$. Then u_n, defined recursively by* (1.1), *is nonincreasing (nondecreasing) in n if $h > 0$ ($h < 0$).*

Proof. We only prove the case $h > 0$. Proceeding by induction,

$$u_1(x) = \int K(x-y) g(y, \phi(y-h))\,dy = \int K(x-h-y) g(y+h, \phi(y))\,dy$$

$$\leqq \int K(x-h-y) g(y, \phi(y))\,dy = u_0(x).$$

Assume that $u_n \leqq u_{n-1}$. Then since Q is order-preserving, we have $u_{n+1} = Q[u_n] \leqq Q[u_{n-1}] = u_n$. Therefore $u_{n+1} \leqq u_n$ for all n and the lemma is proved.

To show uniqueness of clines, we first recall from Remark 3.1 that $g_x \leqq 0$ implies the existence of at least one cline ϕ. Suppose u is another solution of $u = Q[u]$ with $u(-\infty) > a^+$, $u(\infty) < a^-$. Proposition 4.2 with $u_0 = u$ implies that $\phi(x - x_1) - q_0' e^{-\mu n} \leqq u(x) \leqq \phi(x - x_2) + q_0' e^{-\mu n}$ for all n. Letting $n \to \infty$, we have $\phi(x - x_1) \leqq u(x) \leqq \phi(x - x_2)$. Since ϕ is nonincreasing, we may assume that $x_1 < 0$ and $x_2 > 0$.

Let $\underline{u}_0(x) = \phi(x - x_1)$, $\bar{u}_0(x) = \phi(x - x_2)$ and define $\underline{u}_n$, $\bar{u}_n$ recursively by $\underline{u}_{n+1} = Q[\underline{u}_n]$, $\bar{u}_{n+1} = Q[\bar{u}_n]$. Clearly, $\underline{u}_n \leqq u \leqq \bar{u}_n$ and $\underline{u}_n \leqq \phi \leqq \bar{u}_n$ for all n. From Lemma 4.3, $\underline{u}_n$ increases to a nonincreasing function $\underline{u}$ with the properties $\underline{u} \leqq u$, $\underline{u} \leqq \phi$, $\underline{u}(-\infty) = 1$, $\underline{u}(\infty) = 0$ and $\underline{u} = Q[\underline{u}]$. Similarly, $\bar{u}_n$ decreases to a nonincreasing function $\bar{u}$ with the properties $u \leqq \bar{u}$, $\phi \leqq \bar{u}$, $\bar{u}(-\infty) = 1$, $\bar{u}(\infty) = 0$ and $\bar{u} = Q[\bar{u}]$. In order to show that $u = \phi$, it suffices to show that $\underline{u} = \bar{u}$. This follows from Remark 3.2 and the next lemma.

LEMMA 4.4. *Let ϕ_1, ϕ_2 be two nonincreasing solutions of $\phi = Q[\phi]$, $\phi_1 \leqq \phi_2$ which both converge to* 1 *and* 0 *exponentially as $x \to \mp\infty$. Then $\phi_1 \equiv \phi_2$.*

Before we can prove Lemma 4.4, we must establish two lemmas.

From Proposition 4.2, $\phi_2(x+h)\leqq\phi_1(x)$ for some $h>0$. Let h be the infimum of such h. We assume that $h>0$ and derive from it a contradiction. Note that from (viii*) of (1.12), the function $\int K(x-y)[g(y+h,\phi(y+h))-g(y,\phi(y+h))]\,dy$ is not identically zero if $h\neq 0$. Therefore translation of ϕ_i is not a solution of $\phi=Q[\phi]$.

Let $u_\varepsilon(x)=\phi_2(x+h-\varepsilon)$ for $0\leqq\varepsilon\leqq h/2$. According to the definition of h, we have $u_\varepsilon(x)>\phi(x)$ on some interval for sufficiently small $\varepsilon>0$. Write

$$u_\varepsilon(x)=\int K(x-y)g(y,u_\varepsilon(y))\,dy+n_\varepsilon(x),$$

where $n_\varepsilon(x)=\int K(x-y)[g(y+h-\varepsilon,u_\varepsilon(y))-g(y,u_\varepsilon(y))]\,dy$ is nonpositive but not identically zero. Let $\psi_\varepsilon(x)=u_\varepsilon(x)-\phi_1(x)$. Then

$$\psi_\varepsilon(x)=\int K(x-y)h_\varepsilon(y)\psi_\varepsilon(y)\,dy+n_\varepsilon(x), \tag{4.7}$$

where we set

$$h_\varepsilon(x)=\frac{g(x,u_\varepsilon(x))-g(x,\phi_1(x))}{u_\varepsilon(x)-\phi_1(x)}\geqq 0.$$

We shall employ the following notation: $\mathscr{H}=L^2(\mathbb{R})$ with inner product $(\cdot,\cdot)$, $\psi^+=\max\{\psi,0\}$, $K_\varepsilon:\mathscr{H}\to\mathscr{H}$ is the linear operator

$$K_\varepsilon\psi(x)=\int K(x-y)h_\varepsilon(y)\psi(y)\,dy.$$

From Young's inequality, K_ε is bounded. Observe that $\psi_\varepsilon^+\not\equiv 0$ for every $\varepsilon>0$ sufficiently small but $\psi_0^+\equiv 0$. Finally for an operator $A:\mathscr{H}\to\mathscr{H}$, the symbols $\sigma(A)$, $r(A)$, A^* and $\|A\|$ will denote respectively the spectrum of A, spectral radius of A, adjoint of A and operator norm of A.

We state two lemmas and defer their proofs until after we have proved Lemma 4.4.

LEMMA 4.5. *For $\varepsilon>0$ sufficiently small* (i) *K_ε is a positive operator in the sense that $\psi\geqq 0$ implies that $K_\varepsilon\psi\geqq 0$;* (ii) *K_ε is quasi-compact, i.e., there exist operators V_ε and C_ε such that $\|C_\varepsilon\|<1$, V_ε is compact and $K_\varepsilon=C_\varepsilon+V_\varepsilon$,* (iii) $\lim_{\varepsilon\downarrow 0}\|K_\varepsilon-K_0\|=0$.

LEMMA 4.6. $r(K_0)<1$.

Proof of Lemma 4.4. From (4.7), $\psi_\varepsilon\leqq K_\varepsilon\psi_\varepsilon$ so that $\psi_\varepsilon^+\leqq[K_\varepsilon\psi_\varepsilon]^+\leqq K_\varepsilon\psi_\varepsilon^+$. Since the operator K_ε is order-preserving, an inductive argument shows that $K_\varepsilon^n\psi_\varepsilon^+\geqq\psi_\varepsilon^+\geqq 0$ for all n. From our hypotheses, $\psi_\varepsilon\in\mathscr{H}$. Therefore $\|K_\varepsilon^n\psi^+\|_2\geqq\|\psi_\varepsilon^+\|_2$ which implies that $\|K_\varepsilon^n\|^{1/n}\geqq 1$ for all n. Now if we fix n and let $\varepsilon\downarrow 0$, we have from Lemma 4.5, $\|K_0^n\|^{1/n}\geqq 1$. Hence $\lim_{n\to\infty}\|K_0^n\|^{1/n}=r(K_0)\geqq 1$ which contradicts Lemma 4.6. Therefore $h=0$ and $\phi_1\equiv\phi_2$. The proof of Lemma 4.4 is complete.

Proof of Lemma 4.5. K_ε is positive because $K_\varepsilon(x)\geqq 0$ and $h_\varepsilon(x)\geqq 0$ in $\mathbb{R}$. To show that K_ε is quasi-compact, recall from the definition of h_ε that $h_\varepsilon(x)=g_u(x,\theta_\varepsilon)$, where θ_ε is between u_ε and ϕ_1. From hypothesis (x) of (1.12), there exist $\delta>0$, $\theta_1\in(0,1)$ such that $g_u(x,u)\leqq\theta_1<1$ for $u\in[0,\delta]$, $x\geqq N$ or $u\in[1-\delta,1]$, $x\leqq -N$. Since $u_\varepsilon(-\infty)=\phi_1(-\infty)=1$, $u_\varepsilon(\infty)=\phi_1(\infty)=0$. We can choose $a_\varepsilon>N$ such that $|h_\varepsilon(x)|\leqq\theta_1<1$, when $x\in[-a_\varepsilon,a_\varepsilon]^c$.

Define C_ε, $V_\varepsilon:\mathscr{H}\to\mathscr{H}$ by

$$C_\varepsilon\psi(x)=\int K(x-y)\chi_{[-a_\varepsilon,a_\varepsilon]^c}(y)h_\varepsilon(y)\psi(y)\,dy$$

and

$$V_\varepsilon\psi(x)=\int K(x-y)\chi_{[-a_\varepsilon,a_\varepsilon]}(y)h_\varepsilon(y)\psi(y)\,dy.$$

Then $K_\varepsilon=C_\varepsilon+V_\varepsilon$ and V_ε is compact because $\int\int K^2(x-y)\chi^2_{[-a_\varepsilon,a_\varepsilon]}(y)h^2_\varepsilon(y)\,dy\,d\dot{x}$ is finite. For C_ε, we have $\|C_\varepsilon\psi\|_2=\|K*[\chi_{[-a_\varepsilon,a_\varepsilon]^c}h_\varepsilon\psi]\|_2\leqq\theta_1\|K\|_1\|\psi\|_2$. Therefore $\|C_\varepsilon\|\leqq\theta_1<1$ and K_ε is quasi-compact.

Finally it is elementary to show that h_ε converges to h_0 pointwise as $\varepsilon\downarrow 0$ and $\|(K_\varepsilon-K)\psi\|_2=\|K*[h_\varepsilon-h_0]\psi\|_2\leqq\|K\|_2\|[h_\varepsilon-h_0]\psi\|_1\leqq\|K\|_2\|h_\varepsilon-h_0\|_2\|\psi\|_2$. From the fact that $h_\varepsilon(x)=g_u(x,\theta_\varepsilon)$, where θ_ε is between u_ε and ϕ_1, we have

$$|h_\varepsilon(x)-h_0(x)|=|g_{uu}(x,\xi)||\theta_\varepsilon-\theta_0|\leqq\text{const.}\,|\theta_\varepsilon-\theta_0|.$$

But then ϕ_1, ϕ_2 converge to 1 and 0 exponentially as $x\to\mp\infty$. Thus $|\theta_\varepsilon-\theta|$ is dominated by a square integrable function independently of ε. From the dominated convergence theorem, $\lim_{\varepsilon\downarrow 0}\|h_\varepsilon-h_0\|_2=0$. This establishes (iii) and completes the proof of Lemma 4.5.

Proof of Lemma 4.6. From (4.7), we have

$$\psi_0(x)=K_0\psi_0(x)+n_0(x), \tag{4.8}$$

where ψ_0 and n_0 are both nonpositive and not identically zero.

From Lemma 4.5, $K_0^*=C_0^*+V_0^*$. As is well known, $\|C_0^*\|=\|C_0\|$ and V_0^* is compact if and only if V_0 is compact. Therefore K_0^* is also quasi-compact. In fact, K_0^* is the operator $K_0^*\psi(x)=h_0(x)\int K(y-x)\psi(y)\,dy$. Therefore K_0^* is a positive operator.

According to [9, Thm. 4], since K_0^* is a positive operator $r(K_0^*)\in\sigma(K_0^*)$. If $r(K_0^*)<1$, then $r(K_0)=r(K_0^*)<1$ and the lemma is proved. We cannot have $r(K_0^*)\geqq 1$. For if so, $r(K_0^*)\notin\sigma(C_\varepsilon)$ since $\|C_\varepsilon\|<1$. However, K_ε is a perturbation of C_ε by a compact operator. Weyl's lemma says that perturbation by a compact operator can only change the spectrum of an operator by eigenvalues, [8]. Therefore, $r(K_0^*)$ is an eigenvalue of K_0^* and clearly has the largest modulus among the eigenvalues of K_0^*. By [9, Thm. 5, Cor. 1] applied to $r(K_0^*)$, there exists a nonnegative eigenfunction e_0 corresponding to $r(K_0^*)$. That is to say, $r(K_0^*)e_0(x)=h_0(x)\int K(y-x)e_0(y)\,dy\geqq 0$. Using the same idea we used to show $\phi'<0$ at the beginning of the proof of Lemma 4.1, we see that $e_0(x)>0$ in $\mathbb{R}$.

From (4.8), we have $(\psi_0,e_0)=(K_0\psi_0,e_0)+(n_0,e_0)<(K_0\psi_0,e_0)=(\psi_0,K_0^*e_0)=r(K_0^*)(\psi_0,e_0)$. Since $\psi_0\leqq 0$, we have $r(K_0^*)<1$ which is a contradiction to our assumption. Lemma 4.6 is therefore established and so is Theorem 2.2.

5. Proof of Theorem 2.3. The argument given after Lemma 4.3 actually provides a proof for part (i) of Theorem 2.3. Letting $\underline{u}_0(x)=\phi(x-h_1)$ and $\bar{u}_0(x)=\phi(x-h_2)$, we have $\underline{u}_n\leqq u_n\leqq\bar{u}_n$ for all n, and $\underline{u}_n$, $\bar{u}_n$ converge monotonically to the (unique) cline ϕ. With all the properties $\underline{u}_n$, $\bar{u}_n$ and ϕ have, it is an elementary exercise to show that the convergence is uniform in $\mathbb{R}$.

To prove part (ii) of Theorem 2.3, we first observe that $\|u_n'\|_\infty\leqq\|K_a\|_1+p_1+p_2$ so that $\{u_n\}$ is an equicontinuous sequence of functions. By the Arzela–Ascoli theorem, a subsequence $\{u_{n_k}\}$ will converge uniformly on compact subsets of $\mathbb{R}$ to some continuous function U. From Proposition 4.2, $\phi(x-x_1)\leqq U(x)\leqq\phi(x-x_2)$. We may assume that $x_1<0$ and $x_2>0$. Apply part (i) of Theorem 2.3 with U as the initial data. Then

$U_n \equiv Q^n[U]$ converges uniformly to ϕ as $n \to \infty$. The convergence of u_{n_k} to U is also uniform in $\mathbb{R}$, since u_n, ϕ are uniformly close to 1 and 0 near $\mp\infty$ respectively.

We now state a lemma and use it to prove part (ii) of Theorem 2.3. The proof of the lemma will be given at the end of this section.

LEMMA 5.1. *Given* $\varepsilon > 0$, *there exists* $\delta' > 0$ *such that if* $\|v_0 - \phi\|_\infty \leqq \delta'$, *then* $\|v_n - \phi\|_\infty \leqq \varepsilon$ *for all* n.

Let $M = \sup_{\mathbb{R}\times[0,1]} g_u(x,u)$. We have $\|Q^n[u] - Q^n[v]\|_\infty \leqq M^n \|u - v\|_\infty$ for all n and $u, v \in \mathscr{C}$. For any $\varepsilon > 0$, let δ' be chosen as in Lemma 5.1 and let k_1, k_2 be positive integers such that

$$\|U_k - \phi\|_\infty \leqq \frac{\delta'}{2} \quad \text{if } k \geqq k_1, \qquad \|u_{n_k} - U\|_\infty \leqq \frac{\delta'}{2M^{k_1}} \quad \text{if } k \geqq k_2.$$

Then

$$\left\|Q^{k_1}[u_{n_k}] - Q^{k_1}[U]\right\|_\infty \leqq M^{k_1}\|u_{n_k} - U\|_\infty \leqq \frac{\delta'}{2} \quad \text{if } k \geqq k_2.$$

Furthermore,

$$\begin{aligned}\left\|u_{n_k+k_1} - \phi\right\|_\infty &\leqq \left\|Q^{k_1}[u_{n_k}] - Q^{k_1}[U]\right\|_\infty + \left\|Q^{k_1}[U] - \phi\right\|_\infty \\ &\leqq \delta' \quad \text{if } k \geqq k_2.\end{aligned}$$

Now set $k = k_2$ and $v_0 = u_{n_k+k_1}$. From Lemma 5.1, we have $\limsup_{n\to\infty}\|u_n - \phi\| \leqq \varepsilon$. Since $\varepsilon > 0$ is arbitrary, Theorem 2.3 is proved.

Proof of Lemma 5.1. We have to look carefully into the proofs of Proposition 4.2 and Lemma 4.1. Recall from Remark 4.1 that the only requirements on z_0 and q_0 for Lemma 4.1 to hold are $z_0 \leqq 0$ and that $q_0 > 0$ be sufficiently small. Also, in the proof of the left-hand inequality in Proposition 4.2, k, z_0, q_0 have to satisfy the condition $\phi(x - z_0) - q_0 \leqq u_k(x)$ in $\mathbb{R}$. From our hypothesis, $\phi(x) - \delta' \leqq v_0(x)$ in $\mathbb{R}$. Therefore, we set $z_0 = 0$ and $\delta' = q_0 = q_0'$ small enough to obtain the inequality $\phi(x - z_n) - q_0 e^{-\mu n} \leqq v_n(x)$ for all n. Again from the proof of Lemma 4.1, $\lim_{n\to\infty} z_n = x_1 = (\theta - M)q_0\theta_2^{-1}(1 - e^{-\mu})^{-1}$. Now (4.1) is valid with the same θ_1, δ, N if we decrease $q_0 > 0$. Consequently, θ, M, θ_2, μ above are independent of δ' $(= q_0)$ if δ' is sufficiently small. Hence let δ' be so small that $|x_1| \leqq \varepsilon/2\|\phi'\|_\infty$ and $q_0 \leqq \varepsilon/2$. Then

$$\phi(x) - q_0 e^{-\mu n} = \phi(x - x_1) - q_0 e^{-\mu n} + \phi(x) - \phi(x - x_1) \leqq u_n(x) + \frac{\varepsilon}{2}.$$

This implies that $\phi(x) - \varepsilon \leqq u_n(x)$ for all n which is half of Lemma 5.1. The other half may be proved similarly.

6. Proof of Theorem 2.4. Let $T: \mathscr{H} \to \mathscr{H}$ be the bounded linear operator

$$T\psi(x) = \int K(x - y) g_u(y, \phi(y))\psi(y)\, dy.$$

It is easy to see that T is positive and quasi-compact. Furthermore, the proof of Lemma 4.6 can be used to show that $r(T) < 1$. In place of (4.8), we use

$$\phi'(x) = \int K(x - y) g_u(y, \phi(y))\phi'(y)\, dy + \int K(x - y) g_y(y, \phi(y))\, dy,$$

where the last term is nonpositive and not identically zero.

Choose $\lambda \in (r(T),1)$ and $\eta \in \mathscr{H}$ such that $\eta(x)>0$ in $\mathbb{R}$. Then $\lambda w - Tw = \eta$ has a unique solution $w \in \mathscr{H}$. Since $(\lambda - T)^{-1} = \sum_{j=0}^{\infty} T^j/\lambda^{j+1}$, we see that $w \geqq 0$. In fact $\eta > 0$ implies that $w(x)>0$ in $\mathbb{R}$. By adjusting η, we may assume that $\|w\|_\infty = 1$ and that w is sufficiently regular.

Consider the following inequality:

$$\int K(x-y) g_u(y,\phi(y))\,dy \leqq \int_{|y|\geqq N_1} K(x-y) g_u(y,\phi(y))\,dy + \text{const.} \int_{|y|\leqq N_1} K(x-y)\,dy.$$

From condition (xi) of (1.12), $g_u(x,\phi(x)) < \theta_1 < 1$ if $|x| \geqq N_1$ for some large N_1. Therefore, $\int K(x-y) g_u(y,\phi(y))\,dy \leqq \theta_1 < 1$ if $|x| \geqq N_2$. We extend g to $\mathbb{R}\times\mathbb{R}$ so that $g_u(x,u) \geqq 0$ and $M = \frac{1}{2}\sup_{\mathbb{R}\times\mathbb{R}} g_{uu}(x,u)$ is finite.

Choose $\mu > 0$ such that $\lambda < e^{-\mu} < 1$. On the interval $|x| \leqq N_2$, let $w(x) \geqq m > 0$. Define $m_1 = \sup_{\mathbb{R}} \int K(x-y) g_u(y,\phi(y))\,dy$, $\gamma = (e^{-\mu} - \lambda)m/(m_1 - \theta_1)$, $\beta = \gamma(e^{-\mu} - \theta_1)/M(1+\gamma)^2$ and $z_n(x) = \beta(w(x)+\gamma)e^{-\mu n}$ for all n.

We claim that $N[z_n] \leqq z_{n+1}$ for all n where

$$N[z](x) = \int K(x-y)[g(y,\phi(y)+z(y)) - g(y,\phi(y))]\,dy.$$

Write

$$N[z](x) = \int K(x-y) g_u(y,\phi(y)) z(y)\,dy + \int K(x-y) h(y,z(y)) z(y)\,dy,$$

where

$$h(x,z) = \frac{g(x,\phi(x)+z) - g(x,\phi(x))}{z} - g_u(x,\phi(x)).$$

By the mean value theorem, $|h(x,z)z| \leqq M|z|^2$.

To begin we have $Tz_n(x) = \beta e^{-\mu n}[\lambda w(x) - \eta(x) + \gamma \int K(x-y) g_u(y,\phi(y))\,dy]$ and $N_1[z_n](x) = \int K(x-y) h(y,z_n(y)) z_n(y)\,dy$ satisfies the inequality

$$|N_1[z_n](x)| \leqq M \int K(x-y)|z_n(y)|^2\,dy \leqq M\beta^2 e^{-\mu n}(1+\gamma)^2.$$

Hence $N[z_n](x) \leqq \beta e^{-\mu n}[\lambda w(x) + \gamma \int K(x-y) g_u(y,\phi(y))\,dy + \beta M(1+\gamma)^2]$.

If $|x| \geqq N_2$, the term inside the square bracket is bounded above by $e^{-\mu} w(x) + \gamma\theta_1 + \beta M(1+\gamma)^2$ which by the definition of β is equal to $(w(x)+\gamma)e^{-\mu}$. Therefore, $N[z_n](x) \leqq z_{n+1}(x)$ if $|x| \geqq N_2$. On the other hand if $|x| \leqq N_2$, we have

$$\begin{aligned}(\lambda - e^{-\mu}) w(x) + \gamma \int K(x-y) g_u(y,\phi(y))\,dy + \beta M(1+\gamma)^2 - \gamma e^{-\mu} \\ \leqq (\lambda - e^{-\mu}) m + \gamma(m_1 - e^{-\mu}) + \beta M(1+\gamma)^2 \\ = (\lambda - e^{-\mu}) m + \gamma(m_1 - \theta_1) = 0,\end{aligned}$$

and so the term inside the square bracket is bounded above by $e^{-\mu}(w(x)+\gamma)$. Therefore, $N[z_n](x) \leqq z_{n+1}(x)$ if $|x| \leqq N_2$ and our claim is proved.

Finally let $\delta=\beta\gamma$ in the statement of Theorem 2.4 and $v_n=u_n-\phi$ for all n. If $v_0\leqq\delta$, then $v_0\leqq z_0$ since $w>0$. Proceeding inductively, suppose $v_n\leqq z_n$, then

$$v_{n+1}(x)=u_{n+1}(x)-\phi(x)=\int K(x-y)\left[g(y,\phi(y)+v_n(y))-g(y,\phi(y))\right]dy$$

$$\leqq\int K(x-y)\left[g(y,\phi(y)+z_n(y))-g(y,\phi(y))\right]dy$$

$$=N[z_n](x)\leqq z_{n+1}(x).$$

Hence, $u_n(x)-\phi(x)\leqq Ce^{-\mu n}$ for all n where $C=\beta(1+\gamma)$. This proves half of Theorem 2.4.

To show the other half, we first observe that the proof of $N[z_n]\leqq z_{n,+1}$ also shows that $N[-z_n]\geqq -z_{n+1}$ for all n. This part involves no more than changing the sign of some of the terms in the proof of $N[z_n]\leqq z_{n+1}$.

Now suppose $v_0\geqq-\delta\geqq-z_0$. Proceeding inductively as before, assuming that $v_n\geqq-z_n$, we have

$$u_{n+1}(x)-\phi(x)=\int K(x-y)\left[g(y,\phi(y)+v_n(y))-g(y,\phi(y))\right]dy$$

$$\geqq\int K(x-y)\left[g(y,\phi(y)-z_n(y))-g(y,\phi(y))\right]dy$$

$$=N[-z_n](x)\geqq-z_{n+1}(x).$$

Therefore, $\phi(x)-u_n(x)\leqq Ce^{-\mu n}$ for all n. The proof of Theorem 2.4 is now complete.

Note added in proof. Since this paper was accepted, Dr. Odo Diekmann in Amsterdam has informed the author that some of the results in this paper overlap with his paper, *Clines in a discrete time model in population genetics*, Proc. Conference on Models of Biological Growth and Spread, W. Jäger, ed., Lecture Notes in Biomathematics, 38, Springer-Verlag, New York, 1981.

REFERENCES

[1] C. Conley, *An application of Wazewski's method to a nonlinear boundary value problem which arises in population genetics*, J. Math. Biol., 2 (1975), pp. 241–249.

[2] P. Creegan and R. Lui, *Some remarks about the wave speed and travelling wave solutions of a nonlinear integral operator*, J. Math. Biol., 20 (1984), pp. 59–68.

[3] J. Felsenstein, *The theoretical population genetics of variable selection and migration*, Ann. Rev. Genet., 10 (1976), pp. 253–280.

[4] P. C. Fife and L. A. Peletier, *Nonlinear diffusion in population genetics*, Arch. Rat. Mech. Anal., 64 (1977), pp. 93–109.

[5] _____, *Clines induced by variable selection and migration*, Royal Soc. London Proc. B, 214 (1981), pp. 99–123.

[6] R. A. Fisher, *The advance of advantageous genes*, Ann. Eugenics, 7 (1937), pp. 355–369.

[7] J. B. S. Haldane, *The theory of the cline*, J. Genet., 48 (1948), pp. 277–284.

[8] P. R. Halmos, *A Hilbert Space Problem Book*, D. van Nostrand Co., Inc. Princeton, NJ, 1967.

[9] S. Karlin, *Positive operators*, J. Math. Mech., 8 (1959), pp. 907–937.

[10] R. Lui, *A nonlinear integral operator arising from a model in population genetics* I, Monotone initial data, this Journal, 13 (1982), pp. 913–937.

[11] _____, *A nonlinear integral operator arising from a model in population genetics* II, *Initial data with compact support*, this Journal, 13 (1982), pp. 938–953.

[12] _____, *Existence and stability of travelling wave solutions of a nonlinear integral operator*, J. Math. Biol., 16 (1983), pp. 199–220.

[13] R. LUI, *A nonlinear integral operator arising from a model in population genetics* III, *heterozygote inferior case*, this Journal, 16 (1985), pp. 1180–1206.
[14] R. M. MAY, J. A. ENDLER AND R. E. MCMURTRIE, *Gene frequency clines in the presence of selection opposed by gene flow*, Amer. Nat., 109 (1975), pp. 659–676.
[15] T. NAGYLAKI, *Conditions for the existence of clines*, Genetics, 80 (1975), pp. 565–615.
[16] _____, *Clines with variable migration*, Genetics, 83 (1976), pp. 867–886.
[17] _____, *Clines with asymmetric migration*, Genetics, 88 (1978), pp. 813–827.
[18] J. P. PAUWELUSSEN AND L. A. PELETIER, *Clines in the presence of asymmetric migration*, J. Math. Biol., 11 (1981), pp. 207–233.
[19] S. SAWYER, *Results for inhomogeneous selection migration models*, in preparation.
[20] M. SLATKIN, *Gene flow and selection in a cline*, Genetics, 75 (1973), pp. 733–756.
[21] H. F. WEINBERGER, *Long-time behavior of a class of biological models*, this Journal, 13 (1982), pp. 353–396.

SIAM J. MATH. ANAL.
Vol. 17, No. 1, January 1986

016

STABILITY OF A SURFACE DETERMINED FROM MEASURES OF POTENTIAL*

CARLO DOMENICO PAGANI†

Abstract. Let G be a bounded domain in R^3 whose boundary Γ is connected. Let $V_G\sigma$ be the Newtonian potential produced by a mass of density σ distributed over G. We assume a model distribution for σ and consider the problem of finding G from the knowledge of the potential (a) on Γ, (b) on Γ_a, a large spherical surface surrounding G. These problems are unstable and we investigate what kind of supplementary information on the solutions is needed in order to restore stability. We prove that, in case (a), solutions whose H^s norm ($s>3$) (Sobolev norm) is bounded by a given constant constitute a Hölder-stable class; in case (b), a class of Hölder-stable solutions consists of real analytic functions whose derivatives are suitably bounded.

Introduction. Let G be a bounded domain in R^3 whose boundary Γ is connected. Let $V_G\sigma$ be the Newtonian (electrostatic) potential produced by a mass (charge) of density σ distributed over G,

$$V_G\sigma(x)=\int_G \sigma(y)|x-y|^{-1}dy. \tag{0.1}$$

The classical inverse problems for the potential are of two kinds: a) we are given G; find σ by measuring the potential outside of G. b) we assume a model distribution for σ: $\sigma=\tilde{\sigma}$ say; then find G by measuring the potential.

Both problems have been considered by several authors. A common feature to these problems is that they are ill posed in the sense of Hadamard. The most striking aspect of ill-posedness is their instability, namely the impossibility of efficiently determining the solution from the measured data. This aspect of the problem has been analyzed, in the case a), in [7], [12], [3]. Here we wish to consider problem b). Our aim is to find classes of stable solutions. Let us pose precisely the problem. Let G_0 be a reference bounded domain in R^3 whose boundary Γ_0 is connected and smooth: although it is not necessary, we shall assume through this paper that G_0 is a ball, centered at the origin, with radius r_0. Let z: $\Gamma_0\to R$ be a function belonging to the set

$$\mathscr{A}\equiv\{z\in C^2(\Gamma_0);\ |z|\leqq z^*\} \tag{0.2}$$

where z^* is a given constant, $z^*<r_0$. Let us consider the map

$$\begin{aligned}&\Phi_z\colon \Gamma_0\to R^3,\\ &\Gamma_0\ni\omega\to\Phi_z(\omega)=\omega+z(\omega)n_\omega\end{aligned} \tag{0.3}$$

(n_ω is the outward unit normal to Γ_0 at ω) and define Γ_z as the image of Γ_0 given by Φ_z: $\Gamma_z=\Phi_z\Gamma_0$; G_z is the bounded domain whose boundary is Γ_z. We assume that Φ_z induces a diffeomorphism of class C^2 between Γ_z and Γ_0. Finally, let Γ_a be a surface of a large ball, centered at the origin, with radius $r_0+a(a>z^*)$. Let $\tilde{\sigma}$ be a positive smooth function (typically, $\tilde{\sigma}$ is a constant) and consider the potential $V_{G_z}\tilde{\sigma}$, that we shall denote simply by $V_z\tilde{\sigma}$. Let us define $U_z\sigma$ by writing

$$V_z\sigma=V_0\sigma+U_z\sigma. \tag{0.4}$$

*Received by the editors April 19, 1984.
† Department of Mathematics, Politechnic of Milano, 20133 Milano, Italy.

We define now two operators: $A: \mathscr{A} \subset L^2(\Gamma_0) \to L^2(\Gamma_0)$ and $B: \mathscr{A} \subset L^2(\Gamma_0) \to L^2(\Gamma_0)$ as follows (the physical meaning of these definitions is explained below)

$$A[z] \equiv U_z \tilde{\sigma} \circ \Phi_z, \tag{0.5}$$

$$B[z] \equiv U_z \tilde{\sigma} \circ \Phi_a \tag{0.6}$$

(here $\circ$ means composition). Consequently we shall pose two problems:

Problem I (resp. II). Find z: $A[z]=v$ (resp. $B[z]=v$), where v is a given function defined on Γ_0.

We expect that both problems are ill posed, for the image of A is contained in the Sobolev space $H^{3/2}(\Gamma_0)$ and the image of B is a subset of $C^\infty(\Gamma_0)$. Then the maps: $z \to A[z]$ and $z \to B[z]$ are smoothing, so that the inverse maps will not be continuous in L^2-norm. Thus an approximate knowledge of the potential v does not permit one to recover z. We shall investigate which kind of supplementary information on the solutions is needed in order to stabilize the problems. We shall prove, for Problem I, that a subset of $\mathscr{A}$ consisting of those functions z whose H^s-norm ($s > 3$) is bounded by a given constant, is a class of Hölder type stability, i.e., the inverse of the restriction of A to this subset is a Hölder continuous operator. The precise statement is our Theorem 2.7. A similar result is proved for Problem II: a class of Hölder-stable solutions consists now of real analytic functions z whose derivatives are suitably bounded (see Theorem 2.8).

Inverse problems for potentials arise in many fields of applied sciences; in particular Problem II is classical in geophysics: it is the problem of finding the position and the shape of a body G_z of known density $\tilde{\sigma}$ from the measure of the potential generated by this body on the surface Γ_a surrounding G_z. For that problem one can find uniqueness results in [6] (with $\tilde{\sigma}=1$) and in [9] (and several other papers of the same author) (with $\tilde{\sigma}>0$). Results about stability are given in [9] and in [1]. In this last paper, M. M. Lavrent'ev states that solutions of class C^1 with the first derivatives bounded by a given constant are stable: however, the stability is very poor, being of logarithmic type. Our result is in some sense complementary to that one: for we exhibit a stronger stability (namely of Hölder type) for a restricted class of solutions (analytic functions). Problem I, in case the right member of the equation $A[z]v$ is a constant and z is negative (i.e., the unknown surface Γ_z is contained in the ball G_0) has a plain physical interpretation; see Schaeffer [11], who studied the problem (in [11] $\tilde{\sigma}$ is 1 and Γ_0 any C^∞ surface) and proved the existence of a C^∞ solution. Here, for any right member, we prove a stability estimate (hence uniqueness). Existence results for a similar problem are proved also in [13].

The stability of the maps A and B is studied via a lemma (Lemma 2.1) which rests on the knowledge of the spectrum of some linear approximations of A and B. §1 is devoted to the study of these linear approximations; it is written in a self-consistent form, independent on the remaining part of the paper. We study there the classical inverse problem for the single layer potential: to determine the density of a single layer supported by a given surface Γ_z from the knowledge of the potential created by this layer (on Γ_z or on Γ_a).

Section 2 is devoted to our main problems I and II. First (§2.1) we prove the basic lemma which gives us a criterion for Hölder stability of a nonlinear map. Then (§2.2) we prepare the problems I and II in such a way to apply this criterion; in §2.3 the main results (Theorems 2.7 and 2.8) are stated and proved, by taking account of the results of §1.

Remark about the notation. We shall use through this paper the Sobolev spaces $H^s(\Gamma)$ (for the definition and main properties of such spaces see, e.g., [2, Chap. 7]); we will use the notation

$$\|u\|_s \quad \text{for } \|u\|_{H^s(\Gamma_0)}.$$

1. The single layer potential. Let us consider a fixed surface Γ_z, ($z \in \mathscr{A}$) and the single layer potential generated by a density distributed over Γ_z; in the region $\Gamma_0 \times (-z^*, z^*)$ let us use the coordinates (ω, t) to denote points $y = \omega + tn_\omega$, $\omega \in \Gamma_0$, $t \in (-z^*, z^*)$; let $j(\omega, t)$ be the jacobian of the transformation $y \to (\omega, t)$. Now by using surface coordinates on Γ_0, let us consider the potential written in the form (the reasons for this form will appear clear in §2):

$$W_z\rho(x) \equiv \int_{\Gamma_0} \frac{\rho(\omega)\, j(\omega, z(\omega))\, d\omega}{|x - (\omega + z(\omega)n_\omega)|}. \tag{1.1}$$

We shall consider two circumstances: i) the potential is measured on Γ_z; ii) the potential is measured on Γ_a. Consequently we pose

$$M_z[\rho] \equiv W_z\rho \circ \Phi_z, \tag{1.2}$$

$$N_z[\rho] \equiv W_z\rho \circ \Phi_a \tag{1.3}$$

and formulate two problems:

Problem i (resp. ii). Find ρ: $M_z[\rho] = u$ (resp. $N_z[\rho] = u$), where u is a given function defined on Γ_0.[1]

1.1. Problem i): existence, uniqueness and instability. A classical solution to problem i) is given by the following theorem.

THEOREM 1.1. *For every $u \in H^1(\Gamma_0)$ there exists a unique $\rho \in L^2(\Gamma_0)$ satisfying the equation*

$$M_z[\rho] = u. \tag{1.4}$$

Moreover the estimate

$$\|\rho\|_0 \leqq c(z)\|u\|_1, \tag{1.5}$$

holds where $c(z)$ denotes a constant depending only on the surface Γ_z.

Proof. First we prove the assertion in case Γ_z is a spherical surface, Γ_0 say; next we prove for any z the a priori estimate

$$\|\rho\|_0 \leqq c(z)\|M_z[\rho]\|_1. \tag{1.6}$$

Hence the assertion of the theorem follows from standard principles of functional analysis. The first step proves via a spherical harmonics expansion (for a standard reference on spherical harmonics see, e.g., [5]). If we expand ρ:

$$\rho(\omega) = \sum \alpha_{hj} r_0^n S_{nj}(\theta, \varphi)$$

(θ, φ are the geographical coordinates of a point $\omega \in \Gamma_0$; in the sum n goes from 0 to $+\infty$ and j from 1 to $2n+1$), we get

$$M_0[\rho](\omega) = 4\pi \sum r_0^{n+1}(2n+1)^{-1}\alpha_{nj} S_{nj}(\theta, \varphi).$$

[1] If $\tilde{u}(y)$ is the potential measured on Γ_z, we pose $u(\omega) = \tilde{u}(\omega + z(\omega)n_\omega)$; analogously for Problem ii.

Now, if β_{nj} are the coefficients of the expansion of u, (1.4) (with $z=0$) is uniquely solved by choosing

$$\alpha_{nj}=\frac{1}{4\pi r_0}(2n+1)\beta_{nj}.$$

Then we get

$$\|\rho\|_0=\left(\sum r_0^{2n+2}\alpha_{nj}^2\right)^{1/2}=\frac{1}{4\pi}\left(\sum r_0^{2n}(2n+1)^2\beta_{nj}^2\right)^{1/2}\leqq c\|u\|_1.$$

To prove the second step we can use a representation formula for ρ; we know in fact that, if W is the potential due to a single layer of charge of density σ on a surface Γ, then W is continuous across Γ and the normal derivative suffers a jump of magnitude $4\pi\rho$. Now, let w_+ be the solution of the boundary problem

$$\Delta w_+=0 \quad \text{in } G_z,$$
$$w_+=M_z[\rho]\circ\Phi_z^{-1} \quad \text{on } \Gamma_z$$

and w_- be the solution of the corresponding exterior problem. Then $W_z\rho=w_+$ on G_z and w_- on the complement of G_z. Therefore ρ equals the difference in the normal derivatives

$$\rho=\frac{\partial w_+}{\partial n}\circ\Phi_z-\frac{\partial w_-}{\partial n}\circ\Phi_z$$

(at least within a geometric factor), where n is the outward unit normal to Γ_z.

Now, let $z\in C^{2+s}(\Gamma_0)(s\geqq 0)$. Then, we know (cf. e.g. [2, Chap. 2]) that if $M_z[\rho]\in H^{s+1}(\Gamma_0)$ (so the boundary datum in the previous Dirichlet problem belongs to $H^{s+1}(\Gamma_z)$), then $w_+\in H^{s+3/2}(G_z)$ and its normal derivative belongs to $H^s(\Gamma_z)$; analogous statements hold for w_-. Finally we have

$$\|\rho\|_s\leqq c_s(z)\|M_z[\rho]\|_{s+1}. \tag{1.6'}$$

For the sake of completeness notice also that the converse inequality is true: *if* $z\in C^{2+s}(\Gamma_0)$ $(s\geqq 0)$ *operator* M_z *acts continuously from* $H^s(\Gamma_0)$ *to* $H^{s+1}(\Gamma_0)$ *and the estimate holds*:

$$\|M_z[\rho]\|_{s+1}\leqq c_s(z)\|\rho\|_s. \tag{1.7}$$

This assertion is immediately deducible from the results of [4, §5]. Estimate (1.5) shows a dependence of the density ρ on the gradient of the potential. But an error in the measure of the potential prevents us from having any information on its derivatives, hence on the solution. That is what we called instability.

1.2. Problem i): stabilization. As is known in the theory of ill-posed problems, supplementary information on the solutions can help to restore stability. Such information is usually supplied in the form of a priori bounds on the solutions themselves.

LEMMA 1.2. *Let* $z\in C^{(2-\alpha)/(1-\alpha)}(\Gamma_0)$ $(0<\alpha<1)$. *Then every function* $\rho\in H^{\alpha/(1-\alpha)}(\Gamma_0)$ *satisfies the inequality*

$$\|\rho\|_0\leqq c_\alpha(z)\|M_z[\rho]\|_0^{\alpha}\|\rho\|_{\alpha/(1-\alpha)}^{1-\alpha}. \tag{1.8}$$

Let us define the convex set

$$H_\alpha(E)\equiv\left\{\rho\in L^2(\Gamma_0):\ \|\rho\|_{\alpha/(1-\alpha)}\leqq E\right\}, \tag{1.9}$$

where E is a given positive constant. Then we immediately deduce the following theorem.

THEOREM 1.3. *Let* $z \in C^{(2-\alpha)/(1-\alpha)}(\Gamma_0)$; *every pair of solutions* ρ_1, ρ_2, *to problem* i) *corresponding to data* u_1, u_2 *and belonging to the set* $H_\alpha(E)$ *satisfies the inequality*

$$\|\rho_1 - \rho_2\|_0 \leqq c_\alpha(z) E^{1-\alpha} \|u_1 - u_2\|_0^\alpha. \tag{1.10}$$

Expression (1.10) is a stability estimate. It shows that the restriction of the map M_z to the (non-linear) set $H_\alpha(E)$ has a continuous inverse in $L^2(\Gamma_0)$ and the continuity of the inverse map is Hölderian (with exponent α). In such cases we speak of *Hölder stability*.

Proof of Lemma 1.2. The proof follows from: i) the "a priori" estimate (1.6); ii) inequality (1.7), that we shall use with $s = \alpha/(1-\alpha)$; iii) the estimate

$$\|\psi\|_1 \leqq c_\alpha \|\psi\|_{1/(1-\alpha)}^{1-\alpha} \|\psi\|_0^\alpha \tag{1.11}$$

valid for every function $\psi \in H^{1/(1-\alpha)}(\Gamma_0)$; this estimate is easy to prove by using spherical harmonic expansion and the Hölder inequality. Combining i), iii), and ii), one then gets

$$\begin{aligned} \|\rho\|_0 &\leqq c(z) \|M_z[\rho]\|_1 \\ &\leqq c_\alpha(z) \|M_z[\rho]\|_{1/(1-\alpha)}^{1-\alpha} \|M_z[\rho]\|_0^\alpha \\ &\leqq c_\alpha(z) \|\rho\|_{\alpha/(1-\alpha)}^{1-\alpha} \|M_z[\rho]\|_0^\alpha. \end{aligned}$$

The lemma is proved.

1.3. Problem ii): existence, uniqueness and instability. Let Γ be a spherical surface of radius r; let us define a class of real analytic functions on Γ. Let u be a $L^2(\Gamma)$ function whose coefficients, when expanded in spherical harmonics, are β_{nj}. For reals $s \geqq 0$, $\gamma \geqq 1$ consider the linear space spanned by such functions u that

$$\sum (1+n^2)^s \gamma^{2n} r^{2n+2} \beta_{nj}^2 < +\infty. \tag{1.12}$$

Notice that, by writing $\gamma^{2n} = \sum_{0k}^{\infty} (2n \log \gamma)^k / k!$, we get

$$\text{series (1.12)} \leqq \sum_0^\infty {}_k \frac{2^k (\log \gamma)^k}{k!} \|u\|_{H^{k/2+s}(\Gamma)}^2. \tag{1.13}$$

Definition. We call $\mathscr{U}_\gamma^s(\Gamma)$ the Banach space consisting of those functions $u \in L^2(\Gamma)$ satisfying (1.12) and whose norm is given by the square root of the expression appearing in (1.12). Notice that, if $\gamma = 1$, this space can be identified with $H^s(\Gamma)$. In particular we will use the following values for γ: $\gamma_+ = (r_0 + a)/(r_0 + \sup z)$, $\gamma_- = (r_0 + a)/(r_0 + \inf z)$. We will use the notation:

$$|u; \gamma|_s \quad \text{for } \|u\|_{\mathscr{U}_\gamma^s(\Gamma_0)}.$$

Remark. For every $\rho \in H^s(\Gamma_0)$, $s \geqq 0$, *if* $z \in C^{2+s}(\Gamma_0)$, *then* $N_z[\rho] \in \mathscr{U}_{\gamma_+}^{s+1}(\Gamma_0)$ *and*

$$|N_z[\rho]; \gamma_+|_{s+1} \leqq c_s(z) \|\rho\|_s. \tag{1.14}$$

This assertion can be easily checked for spheres (i.e., $z = \text{constant}$); for general surfaces, consider the restrictions of the potential $W_z \rho$ to Γ_0 (φ say) and to Γ_a (ψ say, obviously related to $N_z[\rho]$ via the map Φ_a). From [4, §5] we deduce that $\varphi \in H^{s+1}(\Gamma_0)$ and $\|\varphi\|_{s+1} \leqq c_s(z) \|\rho\|_s$; then we check that $\|\psi\|_{\mathscr{U}_{\gamma_+}^{s+1}(\Gamma_a)} = \|\varphi\|_{s+1}$ and finally that

$$|N_z[\rho]; \gamma_+|_{s+1} = \frac{1}{\gamma_+} \|\psi\|_{\mathscr{U}_{\gamma_+}^{s+1}(\Gamma_a)}.$$

Now, if we take $u \in \mathscr{U}^1_{\gamma_-}(\Gamma_0)$, i.e. $u(\omega)=\sum \alpha_{nj} r_0^n S_{nj}(\theta,\varphi)$ with $\sum(1+n^2)\alpha_{nj}^2 \gamma_-^{2n} r_0^{2n+2} < +\infty$, then the function U:

$$U(x)=\sum \alpha_{nj}\left(\frac{r_0+a}{|x|}\right)^n r_0^n S_{nj}(\theta,\varphi)$$

is harmonic for $|x|>r_0+\inf z$, $U\circ\Phi_a(\omega)=u(\omega)$, and

$$U\circ\Phi_{\inf z}(\omega)=\sum \alpha_{nj}\gamma_-^n r_0^n S_{nj}(\theta,\varphi)$$

belongs to $H^1(\Gamma_0)$. Thus $U\circ\Phi_z$ also belongs to $H^1(\Gamma_0)$ and there exists a unique density $\rho\in L^2(\Gamma_0)$ (Theorem 1.1) which creates a potential whose restriction to Γ_z is $U\circ\Phi_z$; this single layer potential coincides with $U(x)$ in the region exterior to Γ_z. Thus we have proved the following

THEOREM 1.4. *For every $u\in\mathscr{U}^1_{\gamma_-}(\Gamma_0)$, there exists a unique $\rho\in L^2(\Gamma_0)$ satisfying the equation*

$$N_z[\rho]=u. \tag{1.15}$$

Moreover the estimate holds

$$\|\rho\|_0 \leqq c(z)\,|u;\gamma_-|_1. \tag{1.16}$$

Inequality (1.16) *shows that problem* ii), *as was expected, is much more unstable than Problem* i); *nevertheless, it can be stabilized in the same class $H_\alpha(E)$ as before, as we now show by an example.*

1.4. Problem ii): stabilization.

ASSERTION. *Every pair of solutions ρ_1, ρ_2 (corresponding to data u_1, u_2) of the equation $N_0[\rho]=u$ belonging to the set $H_{1/2}(E)$ satisfy the inequality*

$$\|\rho_1-\rho_2\|_0 \leqq 2E\left\{q\left(\|u_1-u_2\|_0^2/4E^2\right)\right\}^{1/2}. \tag{1.17}$$

Here $0\leqq t\to q(t)$ is a certain concave function, increasing from $q(0)=0$ to $q(+\infty)=+\infty$ and exhibiting the following behavior near the origin.

$$q(t)\approx\left(2\log\left(\frac{r_0}{r_0+a}\right)/\log t\right)^2 \quad \text{as } t\approx 0+. \tag{1.18}$$

Inequality (1.17) shows that the restriction of the map N_0 to $H_{1/2}(E)$ does have a continuous inverse in $L^2(\Gamma_0)$, but the continuity of the inverse operator is now very poor, namely of logarithmic type.

Proof of the assertion. The proof follows, via spherical harmonic expansion, from the application of Jensen's inequality for convex functions. Let $0<\lambda\to p(\lambda)$ be the function given by

$$p(\lambda)=\frac{r_0^2}{4}\lambda^2\gamma^{-2/\sqrt{\lambda}} \qquad (\gamma_-=\gamma_+=\gamma). \tag{1.19}$$

One can easily verify that p is convex; notice also that $p(\lambda)$ and $p(\lambda)/\lambda$ are positive increasing.

The following chain of inequalities holds:

$$p\left(\frac{\|\rho\|_0^2}{\|\rho\|_1^2}\right)=p\left(\frac{\Sigma r_0^{2n+2}\alpha_{nj}^2}{\Sigma r_0^{2n+2}(n(n+1)+1)\alpha_{nj}^2}\right)$$

(α_{nj} are the coefficients of the expansion of ρ in spherical harmonics)

$$\leqq\frac{\frac{1}{4}\Sigma r_0^2 r^{-2\sqrt{n(n+1)+1}} r_0^{2n+2}\alpha_{nj}^2}{\Sigma r_0^{2n+2}(n(n+1)+1)\alpha_{nj}^2}$$

(Jensen's inequality)

$$\leqq\frac{\|N_0[\rho]\|_0^2}{\|\rho\|_1^2}.$$

Thus (1.17) follows by taking $q=p^{-1}$.

On the other hand, we are looking for classes of Hölder stable solutions. Then, let us define the convex set

$$K_\alpha(E)\equiv\left\{\rho\in C^\infty(\Gamma_0):\ |N_z[\rho];\gamma_1|_{1/(1-\alpha)}\leqq E\right\} \tag{1.20}$$

where $\gamma_1=\gamma_-^{1/1-\alpha}$. Notice that, because $\gamma_1>\gamma_-\geqq\gamma_+$, form $N_z[\rho]$ belong to some $\mathscr{U}_{\gamma_1}^{s+1}(\Gamma_0)$ one requires that ρ and z belong to some class of analytic functions; e.g., if z is constant, ρ must belong to $\mathscr{U}_{\gamma_1/\gamma_+}^{s}(\Gamma_0)$.

THEOREM 1.5. *Every pair of solutions ρ_1, ρ_2 to problem* ii) *corresponding to data u_1, u_2 and belonging to the set $K_\alpha(E)$ satisfy the inequality*

$$\|\rho_1-\rho_2\|_0\leqq c_\alpha(z)E^{1-\alpha}\|u_1-u_2\|_0^\alpha. \tag{1.21}$$

Proof of the theorem. The proof follows, analogously to the proof of Lemma 1.2, from inequality (1.16) and from the following:

$$|\psi;\gamma_-|_1\leqq c_\alpha|\psi;\gamma_1|_{1/(1-\alpha)}^{1-\alpha}\|\psi\|_0^\alpha. \tag{1.22}$$

This one proves by using spherical harmonics and Hölder's inequality. Combining (1.16) and (1.22), one gets

$$\begin{aligned}\|\rho\|_0&\leqq c(z)|N_z[\rho];\gamma_-|_1\\&\leqq c_\alpha(z)|N_z[\rho];\gamma_1|_{1/(1-\alpha)}^{1-\alpha}\|N_z[\rho]\|_0^\alpha,\end{aligned}$$

from which (1.21) immediately follows.

2. The nonlinear Problems I and II.

2.1. A criterion for Hölder stability.

Assumptions. a) X, Y are Banach spaces, x_1 is an element of X; for every $x_2\in X$ belonging to some neighborhood of x_1 we write $x_2=x_1+h$.

b) Let C: $X\to Y$ be a map continuously differentiable in a neighborhood of x_1 and such that

$$\|C[x_1+h]-C[x_1]-C'(x_1)[h]\|_Y\leqq c(x_1)\|h\|_Y^{1+\varepsilon} \tag{2.1}$$

for some $\varepsilon > 0$. c) The linear map $C'(x_1)$: $X \to Y$ is not invertible; but there exists a linear operator S: $X \to X$ such that, if we define the set

$$\mathscr{H}(E) \equiv \{ x \in X; \|S[x]\|_X \leqq E \}, \tag{2.2}$$

the estimate

$$\|h\|_X \leqq c(x_1, E, \gamma) \|C'(x_1)[h]\|_Y^{\gamma} \tag{2.3}$$

is valid for some γ: $0 < \gamma < 1$ and every $x_2 \in X$ with $h = x_2 - x_1 \in \mathscr{H}(E)$.

LEMMA 2.1. *Assuming that* a), b), c) *hold, the following assertion is valid: if*

$$(1+\varepsilon)\gamma > 1, \tag{2.4}$$

then there exist numbers $R = R(x_1, E, \gamma)$ *and* $c = c(x_1, E, \gamma)$ *such that the estimate*

$$\|h\|_X \leqq c \|C[x_1 + h] - C[x_1]\|_Y^{\gamma} \tag{2.5}$$

holds for every $h \in \mathscr{H}(E)$ *with* $\|h\|_X \leqq R$.

As a consequence, if x_1, x_2 are solutions to the equation $C[x] = f$ with $f = f_1$, f_2 respectively (and all the previous hypotheses hold), we get, by linearizing around x_1, the stability estimate

$$\|x_2 - x_1\|_X \leqq c(x_1, E, \gamma) \|f_2 - f_1\|_Y^{\gamma}. \tag{2.6}$$

We emphasize that the determination of the stability class $\mathscr{H}(E)$ depends only on the linear approximation C' of C, provided that this class is of Hölder type (with suitable Hölder exponent). A counterexample to the lemma is given in [8]; it shows that, if hypothesis (2.4) is not satisfied, then a bound (on the solutions) which is sufficient to stabilize the linear problem is completely inadequate to stabilize the nonlinear operator.

Proof of Lemma 2.1. Let us consider the identity

$$C'(x_1)[h] = C[x_1 + h] - C[x_1] - (C[x_1 + h] - C[x_1] - C'(x_1)[h]).$$

Now we apply first inequality (2.3), then the triangular inequality, finally (2.1); we get:

$$\|h\|_X \leqq c(x_1, E, \gamma) \{ \|C[x_1 + h] - C[x_1]\|_Y + \|C[x_1 + h] - C[x_1] - C'(x_1)[h]\|_Y \}^{\gamma}$$

$$\leqq c(x_1, E, \gamma) \{ \|C[x_1 + h] - C[x_1]\|_Y^{\gamma} + \|h\|_X^{\gamma(1+\varepsilon)} \}.$$

Because of (2.4) the last term at the right of the previous inequality is dominated by the terms at the left when $\|h\|_X \to 0$. That proves the lemma.

2.2. Linear approximations of *A* and *B*. Let us recall the definition of the potential $U_z\tilde{\sigma}$, given in (0.4), and write it by using surface coordinates

$$U_z\tilde{\sigma}(x) = \int_{\Gamma_0} d\omega \int_0^{z(\omega)} \frac{\tilde{\sigma}(\omega + tn_\omega) j(\omega, t)\, dt}{|x - (\omega + tn_\omega)|}. \tag{2.7}$$

We recall that the given function $\tilde{\sigma}$ is smooth (C^∞ say) and strictly positive, so that the product $\tilde{\sigma}(\omega + tn_\omega) j(\omega, t)$ is also smooth and there exist positive constants c_1, c_2 such

that

$$c_1 \leqq |\tilde{\sigma}(\omega + tn_\omega)\, j(\omega, t)| \leqq c_2. \tag{2.8}$$

In what follows $\tilde{\sigma}$ is considered as fixed and perfectly known and we will ignore the dependence of the relevant quantities of $\tilde{\sigma}$. As we said in the introduction, the map A, defined in (0.5), with domain in $\mathscr{A}$, has values in $H^{3/2}(\Gamma_0)$; while the map B, defined in (0.6), has values in $C^\infty(\Gamma_0)$. The next two lemmas describe a linear approximation of A and B.

LEMMA 2.2. *The map* A: $\mathscr{A} \subset L^2(\Gamma_0) \to L^2(\Gamma_0)$ *is differentiable at the origin; its Fréchet derivative,* $A'(0)$ *say, is given by*

$$A'(0)[\rho] = M_0[\tilde{\rho}\rho] \tag{2.9}$$

(*here is* $\tilde{\rho}(\omega) = \tilde{\sigma}(\omega)$, *see Lemma* 2.3). *Moreover the estimates hold*:

$$\|A[z]\|_0 \leqq c(z^*)\|z\|_0 \tag{2.10}$$

for every $z \in \mathscr{A}$, *and*

$$\|A[\rho] - M_0[\tilde{\rho}\rho]\|_0 \leqq c\|\rho\|_0^{1+\varepsilon} \tag{2.11}$$

for every ε: $0 < \varepsilon < \frac{1}{3}$. *The constant appearing in* (2.11) *depends on* z^* *and on the* $W^{1,\infty}(\Gamma_0)$ *norm of* ρ.

LEMMA 2.3. *The map* B: $\mathscr{A} \subset L^2(\Gamma_0) \to L^2(\Gamma_0)$ *is continuously Fréchet differentiable at any point* $\bar{z} \in \mathscr{A}$; *its Fréchet derivative,* $B'(\bar{z})$ *say, is given by*

$$B'(\bar{z})[\rho] = N_{\bar{z}}[\tilde{\rho}\rho], \tag{2.12}$$

where $\tilde{\rho}(\omega) = \tilde{\sigma}(\omega + \bar{z}(\omega)n_\omega)$; *the estimates*

$$\|B[z]\|_0 \leqq c(z^*)\|z\|_0, \tag{2.13}$$

$$\|B[\bar{z}+\rho] - B[\bar{z}] - N_{\bar{z}}[\tilde{\rho}\rho]\|_0 \leqq c\|\rho\|_0^{1+\varepsilon} \tag{2.14}$$

hold for every ε: $0 < \varepsilon < \frac{1}{3}$. *The constant appearing in* (2.14) *depends on* $z^*, \bar{z}$, *and on the* $W^{1,\infty}(\Gamma_0)$ *norm of* ρ.

Remark. For the operator A we can prove a more complete result, as we did for B: *The map* A *is continuously Fréchet differentiable at any point* $\bar{z} \in \mathscr{A}$; *we have*

$$A'(\bar{z})[\rho] = M_z[\tilde{\rho}\rho] + g_{\bar{z}}\rho \tag{2.9'}$$

where $g_{\bar{z}}$: $\Gamma_0 \to R$ *is the function so defined*

$$g_{\bar{z}}(\omega) = -\int_{R_{\bar{z}}} \frac{\langle \omega + \bar{z}(\omega)n_\omega - y, n_\omega \rangle \tilde{\sigma}(y)\, dy}{|\omega + \bar{z}(\omega)n_\omega - y|^3} \tag{2.15}$$

and $R_{\bar{z}}$ *is the symmetric difference between* G_0 *and* $G_{\bar{z}}$. *Moreover it holds an estimate quite analogous to* (2.11).

The proof of this assertion is rather lengthy; we refrain from writing it out, and observe that in the next section we will use only the partial result stated in Lemma 2.2.

The next three lemmas are technical and will be useful in the proofs of Lemma 2.2 and 2.3.

LEMMA 2.4. *Let* $\omega, \omega' \in \Gamma_0$, t, t', t_1, t_2, δ *reals*, $0 \leqq \delta \leqq 2$, $|t|, |t_1|, |t_2| < z^*$. *The following inequalities hold*:

$$|\omega - \omega' + tn_\omega - t'n_{\omega'}| \geqq c|\omega - \omega'|, \tag{2.16}$$

$$|\omega - \omega' + tn_\omega - t'n_{\omega'}| \geqq c|t - t'|, \tag{2.17}$$

$$|\omega - \omega' + tn_\omega - t'n_{\omega'}|^2 \geqq c|\omega - \omega'|^\delta |t - t'|^{2-\delta} \tag{2.18}$$

$$|\omega - \omega' + t'n_{\omega'} - t_1 n_\omega|^{-1} - |\omega - \omega' + t'n_{\omega'} - t_2 n_\omega|^{-1} \leqq c|t_1 - t_2|^{\delta - 1} |\omega - \omega'|^{-\delta} \tag{2.19}$$

for some positive constants c, *possibly depending on* z^*.

LEMMA 2.5. *For every real valued functions* $\rho, \sigma \in L^2(\Gamma_0)$ *and a real number* $\delta < 2$ *one gets*

$$\int_{\Gamma_0} d\omega' \left(\int_{\Gamma_0} \frac{\rho(\omega)^2 d\omega}{|\omega - \omega'|^\delta} \right)^2 \leqq c \int_{\Gamma_0} \rho(\omega)^4 d\omega, \tag{2.20}$$

$$\int_{\Gamma_0} \sigma(\omega')^2 d\omega' \left(\int_{\Gamma_0} \frac{\rho(\omega) d\omega}{|\omega - \omega'|^\delta} \right)^2 \leqq c \left(\int_{\Gamma_0} \sigma(\omega)^4 d\omega \right)^{1/2} \left(\int_{\Gamma_0} \rho(\omega)^4 d\omega \right)^{1/2}. \tag{2.20'}$$

LEMMA 2.6. *For every real valued function* $\rho \in C^1(\Gamma_0)$ *and positive* $\delta \leqq 3$ *one gets*

$$\int_{\Gamma_0} |\rho(\omega)|^{2\delta} d\omega \leqq c\big(\|\rho\|_{W^{1,\infty}(\Gamma_0)}\big) \left(\int_{\Gamma_0} \rho(\omega)^2 d\omega \right)^{2\delta/3}. \tag{2.21}$$

Proof of Lemma 2.3. First we prove (2.13). We have

$$|B[z](\omega')| = \left| \int_{\Gamma_0} d\omega \int_0^{z(\omega)} \frac{\tilde{\sigma}(\omega + tn_\omega) j(\omega, t)\, dt}{|\omega' + an_{\omega'} - (\omega + tn_\omega)|} \right|$$

$$\leqq c(z^*) \int_{\Gamma_0} \frac{|z(\omega)|\, d\omega}{|\omega - \omega'|}.$$

Here we have used (2.8) and (2.16). Then (2.13) follows from known properties of the single layer potential.

To justify (2.12) and prove (2.14) it is sufficient, because of the definition of B, to estimate the difference

$$U_{\bar{z}+\rho}\tilde{\sigma}(x) - U_{\bar{z}}\tilde{\sigma}(x) - W_{\bar{z}}\rho\tilde{\rho}(x). \tag{2.22}$$

We write it as follows:

$$\int_{\Gamma_0} d\omega \int_{\bar{z}(\omega)}^{\bar{z}(\omega)+\rho(\omega)} dt \left[\frac{\tilde{\sigma}(\omega + tn_\omega) j(\omega, t)}{|x - (\omega + tn_\omega)|} - \frac{\tilde{\rho}(\omega) j(\omega, \bar{z}(\omega))}{|x - (\omega + \bar{z}(\omega) n_\omega)|} \right]$$

$$= \int_{\Gamma_0} d\omega \int_{\bar{z}(\omega)}^{\bar{z}(\omega)+\rho(\omega)} dt \frac{\tilde{\sigma}(\omega + tn_\omega) j(\omega, t) - \tilde{\rho}(\omega) j(\omega, \bar{z}(\omega))}{|x - (\omega + tn_\omega)|}$$

$$+ \int_{\Gamma_0} d\omega \int_{\bar{z}(\omega)}^{\bar{z}(\omega)+\rho(\omega)} dt \tilde{\rho}(\omega) j(\omega, \bar{z}(\omega)) \Big[|x - (\omega + tn_\omega)|^{-1} - |x - (\omega + \bar{z}(\omega) n_\omega)|^{-1} \Big].$$

In the last equality the integrand of the first integral is estimated by $|t-\bar{z}(\omega)||\omega-\omega'|^{-1}$, having put $x=\omega'+an_{\omega'}$, $\omega'\in\Gamma_0$; we took account of (2.16) and the smoothness of $\tilde{\sigma}$ and j. The integrand of the second integral is estimated by $|t-\bar{z}(\omega)|^{\delta-1}|\omega-\omega'|^{-\delta}$, $(0<\delta<2)$, thanks to (2.19). Then we get

$$
\begin{aligned}
&|\text{expression (2.22)}| \\
&\leqq c(z^*)\left[\int_{\Gamma_0} d\omega \int_{\bar{z}(\omega)}^{\bar{z}(\omega)+\rho(\omega)} dt \frac{|t-\bar{z}(\omega)|}{|\omega-\omega'|}+\int_{\Gamma_0} d\omega \int_{\bar{z}(\omega)}^{\bar{z}(\omega)+\rho(\omega)} dt \frac{|t-\bar{z}(\omega)|^{\delta-1}}{|\omega-\omega'|^{\delta}}\right] \\
&\leqq c(z^*)\left[\int_{\Gamma_0} \frac{\rho(\omega)^2 d\omega}{|\omega-\omega'|}+\int_{\Gamma_0} \frac{|\rho(\omega)|^{\delta} d\omega}{|\omega-\omega'|^{\delta}}\right].
\end{aligned}
$$

Now the $L^2(\Gamma_0)$ norm of the first integral (in the last inequality) is plainly estimated by $\|\rho^2\|_0$, while the $L^2(\Gamma_0)$ norm of the second integral, thanks to Lemma (2.5), is estimated by $\||\rho|^{\delta}\|_0$. So we got, for $0<\delta<2$,

$$
\|B[\bar{z}+\rho]-B[\bar{z}]-N_{\bar{z}}[\tilde{\rho}\rho]\|_0 \leqq c(z^*)\left\||\rho|^{\delta}\right\|_0.
$$

Now (2.14) follows from Lemma 2.6. Lemma 2.3 is proved.

Proof of Lemma 2.2. Inequality (2.10) is proved in a quite analogous manner as (2.13). To justify (2.9) and prove (2.11) we have to estimate the difference $A[\rho]-M_0[\tilde{\rho}\rho]$, that we write as follows:

$$
\left[U_\rho\tilde{\sigma}\circ\Phi_\rho-U_\rho\tilde{\sigma}\circ\Phi_0\right]+\left[\left(U_\rho\tilde{\sigma}-W_0\tilde{\rho}\rho\right)\circ\Phi_0\right].
$$

The expression in the second bracket handles exactly as we did before in the proof of Lemma 2.3; then its $L^2(\Gamma_0)$ norm is estimated by $\|\rho\|_0^{1+\varepsilon}$ with $\varepsilon<\frac{1}{3}$. The expression in the first bracket is

$$
\int_{\Gamma_0} d\omega \int_0^{\rho(\omega)} dtj(\omega,t)\tilde{\sigma}(\omega+tn_\omega)\left[|\omega'+\rho(\omega')n_{\omega'}-(\omega+tn_\omega)|^{-1}-|\omega'-(\omega+tn_\omega)|^{-1}\right].
$$

The integrand is estimated by $|\rho(\omega')|^{\delta-1}|\omega-\omega'|^{-\delta}$, thanks to (2.19). Then all the expression is estimated by

$$
|\rho(\omega')|^{\delta-1}\int_{\Gamma_0}\frac{|\rho(\omega)|\,d\omega}{|\omega-\omega'|^{\delta}}
$$

and its $L^2(\Gamma_0)$ norm by $\||\rho|^{\delta}\|_0$, because of (2.20′). Then (2.11) follows by applying Lemma 2.6. Lemma 2.2 is proved.

Proof of Lemma 2.4. Inequalities (2.16) and (2.17) are almost trivial; (2.18) is a simple consequence of the first two; (2.19) is derived from (2.18) by observing that

$$
\frac{d}{dt}|\omega'+t'n_{\omega'}-(\omega+tn_\omega)|^{-1}\leqq|\omega'+t'n_{\omega'}-(\omega+tn_\omega)|^{-2}
$$

and applying the fundamental theorem of calculus.

Proof of Lemma 2.5. From the Schwarz inequality, by taking $\delta-1<\alpha<1$, we have

$$
\text{(2.23)}\qquad \left(\int_{\Gamma_0}\frac{\rho(\omega)^2 d\omega}{|\omega-\omega'|^{\delta}}\right)^2\leqq\int_{\Gamma_0}\frac{d\omega}{|\omega-\omega'|^{2(\delta-\alpha)}}\int_{\Gamma_0}\frac{\rho(\omega)^4 d\omega}{|\omega-\omega'|^{2\alpha}}=c\int_{\Gamma_0}\frac{\rho(\omega)^4 d\omega}{|\omega-\omega'|^{2\alpha}}.
$$

Then

$$\int_{\Gamma_0} d\omega' \left(\int_{\Gamma_0} \frac{\rho(\omega)\, d\omega}{|\omega - \omega'|^{\delta}} \right)^2 \leqq c \int_{\Gamma_0} \rho(\omega)^4 d\omega \int_{\Gamma_0} \frac{d\omega'}{|\omega - \omega'|^{2\alpha}} = c \int_{\Gamma_0} \rho(\omega)^4 d\omega.$$

To prove (2.20′) we proceed analogously, by first applying the Schwarz inequality

$$\int_{\Gamma_0} \sigma(\omega')^2 d\omega' \left(\int_{\Gamma_0} \frac{\rho(\omega)\, d\omega}{|\omega - \omega'|^{\delta}} \right)^2 \leqq \left(\int_{\Gamma_0} \sigma(\omega')^4 d\omega' \right)^{1/2} \left(\int_{\Gamma_0} d\omega' \left(\int_{\Gamma_0} \frac{\rho(\omega)\, d\omega}{|\omega - \omega'|^{\delta}} \right)^4 \right)^{1/2};$$

then, by applying repeatedly (2.23), we get

$$\int_{\Gamma_0} d\omega' \left(\int_{\Gamma_0} \frac{\rho(\omega)\, d\omega}{|\omega - \omega'|^{\delta}} \right)^4 \leqq \int_{\Gamma_0} \rho(\omega)^4 d\omega.$$

Lemma 2.5 is proved.

Proof of Lemma 2.6. In the Sobolev inequality

$$\left(\int_{\Gamma_0} |u(\omega)|^p d\omega \right)^{1/p} \leqq c \int_{\Gamma_0} \left(|u(\omega)| + |Du(\omega)| \right) d\omega,$$

which holds with $p \leqq 2$, take $u = |\rho|^3$; then choose $p = 2\delta/3$; so we get

$$\int_{\Gamma_0} |\rho(\omega)|^{2\delta} d\omega \leqq c \left(\max|\rho| + 3 \max|D\rho| \right)^{2\delta/3} \left(\int_{\Gamma_0} \rho(\omega)^2 d\omega \right)^{2\delta/3}.$$

Lemma 2.6 is proved.

2.3. Stability results for Problems I and II. Now we are in the position to draw some conclusions about the stability of the solutions of Problems I and II, for the results of §1 give us information about the Hölder stability of the linear operators A' and B'. That is true for $B'(z)$, which substantially coincides with N_z, while $A'(z)$, as we noted in the remark after Lemma 2.3, differs from M_z by the perturbation term g_z; so the linearized equation $A'(z)[\rho] = f$ is an integral equation of the second kind. This equation presents problems of instability analogous to those exhibited by the first kind equation: $M_z[\rho] = f$ discussed in §1; for the term g_z is actually small, and it can vanish (if $z \equiv 0$, then, as we saw, $g_z \equiv 0$). Now, using the results of §1 directly, we will take, for Problem I, $\bar{z} = 0$; thus we will get a stability result for solutions of Problem I which are close to the origin.

THEOREM 2.7. *Let $v \in L^2(\Gamma_0)$ be a given function and $z \in \mathscr{A}$ a solution of the equation $A[z] = v$ belonging to the set $H_\alpha(E)$ with $\alpha > \frac{3}{4}$. Then there exist numbers $R = R(\alpha, E)$ and $c = c(\alpha)$ such that the estimate*

$$\|z\|_0 \leqq c E^{1-\alpha} \|v\|_0^{\alpha} \tag{2.24}$$

is valid for every z with $\|z\|_0 \leqq R$.

Proof. We have only to verify that the hypotheses of Lemma 2.1 are fulfilled. We apply Lemma 2.1 by taking $X = Y = L^2(\Gamma_0)$ and $x_1 = 0$. Inequality (2.1) is satisfied with $\varepsilon < \frac{1}{3}$ (cf. inequality (2.11)). The stability class $\mathscr{K}(E)$ is the set $H_\alpha(E)$ defined in (1.9) and inequality (2.3) is satisfied with $\gamma = \alpha$ (cf. inequality (1.10)); notice that the dependence of the constant appearing in (2.3) on E is explicit in (1.10). Then the hypothesis

$(1+\varepsilon)\gamma > 1$ can be satisfied with $\alpha > \frac{3}{4}$; so (2.24) follows. Notice that the dependence of the constants on $\|z\|_{W^{1,\infty}(\Gamma_0)}$ (coming from (2.11)) has disappeared, because the stability constraint: $\|z\|_{\alpha/1-\alpha} \leqq E$ with $\alpha > \frac{3}{4}$ also bounds the $W^{1,\infty}$ norm of z by the same constant E. A more complete stability result can be proved for Problem II.

THEOREM 2.8. *Let $v_1, v_2 \in L^2(\Gamma_0)$ be given functions and $z_1, z_2 \in \mathscr{A}$ be the corresponding solutions of the equation $B[z]=v$. If $z_2 - z_1 \in K_\alpha(E)$ with $\alpha > \frac{3}{4}$ then there exist numbers $R = R(\alpha, E, z_1)$ and $c = c(\alpha, z_1)$ such that the estimate*

$$\|z_2 - z_1\|_0 \leqq cE^{1-\alpha}\|v_2 - v_1\|_0^{\alpha} \tag{2.25}$$

is valid for $\|z_2 - z_1\|_0 \leqq R$.

The proof follows from Lemma 2.1 in a way analogous to the previous theorem.

REFERENCES

[1] M. M. LAVRENT'EV, *Some Improperly Posed Problems of Mathematical Physics*, Springer Tracts in Natural Philosophy, 11, Springer-Verlag, Berlin, 1967.

[2] J. L. LIONS AND E. MAGENES, *Problèmes aux limites non homogènes et applications*, Dunod, Paris, 1968.

[3] A. LORENZI AND C. D. PAGANI, *An inverse problem in potential theory*, Ann. Mat. Pura e Appl., 129 (1981), pp. 281–303.

[4] C. MIRANDA, *Sulle proprietà di regolarità di certe trasformazioni integrali*, Acc. Naz. Lincei, Memorie Sc., serie VIII, vol. VII, sez. I, 9 (1965), pp. 303–336.

[5] C. MÜLLER, *Spherical Harmonics*, Lecture Notes in Mathematics 17, Springer-Verlag, Berlin, 1966.

[6] P. S. NOVIKOV, *Sur le problème inverse du potentiel*, Dokl. Akad. Nauk SSSR, 18 (1938), pp. 165–168.

[7] C. D. PAGANI, *Inverse problems for the volume potential*, Portugaliae Mathematica, 41 (1982).

[8] _____, *Questioni di stabilità per problemi inversi*, Rend. Sem. Mat. Fis. Milano, to appear.

[9] A. I. PRILEPKO, *On the uniqueness of solutions to inverse problems of Newtonian potential*, Differencial'nye Uravnenija, 2 (1966), pp. 107–124.

[10] I. M. RAPOPORT, *On stability in the inverse potential problem*, Dokl. Akad. Nauk SSSR, 31 (1941), pp. 302–304.

[11] D. G. SCHAEFFER, *The capacitor problem*, Indiana Univ. Math. J., 24 (1975), pp. 1143–1167.

[12] N. WECK, *Inverse Probleme der Potentialtheorie*, Appl. Anal., 2 (1972), pp. 195–238.

[13] C. MADERNA AND C. D. PAGANI AND S. SALSA, *Existence results in an inverse problem of potential theory*, *preprint*.

SIAM J. MATH. ANAL.
Vol. 17, No. 1, January 1986

ABSOLUTELY CONTINUOUS SPECTRA OF SECOND ORDER DIFFERENTIAL OPERATORS WITH SHORT AND LONG RANGE POTENTIALS*

D. B. HINTON† AND J. K. SHAW‡

Abstract. For ordinary second order differential operators with one or two singular endpoints, the problem is considered of determining when a continuous spectrum is absolutely continuous or is of class $C^{(1)}$. Operators are considered which have a smooth part plus perturbation terms. The perturbation terms considered are short and long range type potentials and also potentials of a highly oscillatory character. The absolutely continuous spectrum found is either a ray $[\lambda_0, \infty)$ or a whole line $(-\infty, \infty)$. For a certain class of equations, the spectrum is also found to be bounded below. The theory developed is applied to the energy operator of the hydrogen atom.

1. Introduction. As is well known, singular boundary value problems often have continuous spectra. It is of interest in applications to know if this spectrum is absolutely continuous. We consider here a class of such problems for second order ordinary differential equations in both the half line and whole line cases. One version of the problem may be described as follows. Let

$$L(y) = w^{-1}\{-(py')' + qy\} = \lambda y, \qquad a \leqq x < \infty, \tag{1.1}$$

$$\sin\alpha\, y(a) + \cos\alpha\,(py')(a) = 0 \tag{1.2}$$

be an eigenvalue problem such that L is in the limit point case at infinity. Let $\psi(x,\lambda)$ be the solution of (1.1) defined by $\psi(a,\lambda) = \cos\alpha$, $(p\psi')(a,\lambda) = -\sin\alpha$. Then there is a nondecreasing function ρ on $(-\infty, \infty)$ such that the equations

$$g(\lambda) = \int_a^\infty f(x) w(x) \psi(x,\lambda)\, dx,$$

$$f(x) = \int_{-\infty}^\infty g(\lambda)\psi(x,\lambda)\, d\rho(\lambda),$$

define a linear isometry between the Hilbert spaces

$$\mathscr{L}_w^2(a,\infty) = \left\{ f \,\middle|\, \int_a^\infty w|f|^2\, dx < \infty \right\},$$

$$\mathscr{L}^2(-\infty,\infty) = \left\{ g \,\middle|\, \int_{-\infty}^\infty |g|^2\, d\rho < \infty \right\}.$$

We refer the reader to [7. Chap. 9]. The above isometry is a unitary transformation which takes the self-adjoint operator T associated with (1.1)–(1.2) to the self-adjoint transformation in $\mathscr{L}_\rho^2(-\infty,\infty)$ which is multiplication by the independent variable λ. We may then characterize the spectrum of T, $\sigma(T)$, as the points of increase of ρ. Further the eigenvalues of T, $D\sigma(T)$, are the jumps of ρ, and the essential spectrum of T, $E\sigma(T)$, is the set of limit points of $\sigma(T)$. We define the continuous spectrum $C\sigma(T)$

*Received by the editors July 12, 1983, and in final revised form April 19, 1984.
† Department of Mathematics, University of Tennessee, Knoxville, Tennessee 37996-1300.
‡ Department of Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061-4097.

as $E\sigma(T)\backslash D\sigma(T)$. The problem we consider is to find conditions on the coefficients of (1.1) so that if $[a,b]\subset C\sigma(T)$, then $[a,b]$ is an absolutely continuous spectrum, i.e., ρ is absolutely continuous on $[a,b]$. Our conditions will in fact imply ρ is of class $C^{(1)}$.

In order to consider equations like (1.1), we define below a generalization of the equation,

$$y''+\left[\lambda+v_1(x)+v_2(x)+v_3(z)\right]y=0, \qquad a\leqq x<\infty, \tag{1.3}$$

which will include (1.1) after a change of independent and dependent variables. In (1.3) regard v_1 as absolutely continuous with $\int_a^\infty |v_1'|<\infty$, v_2 as having a small integral, and v_3 as being small in the sense of $\int_a^\infty |v_3|<\infty$. Such v_1, v_3 are called long and short range potentials, respectively. An example of the type of equation we wish to consider is,

$$y''+\left[\lambda+(x+1)^{-1}+x\sin x^4+e^{-x}x^{-1/2}\right]y=0, \qquad 0\leqq x<\infty.$$

This work is motivated by recent work of S. Itatsu and H. Kaneta [10], as well as by classical work of E. C. Titchmarsh [16, pp. 116–128]. In [16], the case $v_1=v_2\equiv 0$ is considered as well as certain equations in which a change of variable is made. The change of independent variable is nonreal and places restrictions on the coefficients. In particular, perturbation terms are not allowed. We are able to avoid many of these restrictions by making the change of independent variable real. This however, may make the terms v_1, v_2, and v_3 λ-dependent. In [10], (1.3) is considered with $v_2\equiv 0$, and the authors use the method of singular integral equations to obtain asymptotics for (1.3). We find it more advantageous to develop the theory along the lines of Titchmarsh.

A different approach to showing continuous spectrum is absolutely continuous has been taken by J. Weidmann [18] and J. Walter [17]. They approximate the singular problem by a sequence of regular problems and apply estimates on the number of eigenvalues lying in a given interval. It does not appear that this method can be used to show additional smoothness properties of ρ, e.g., to show ρ is of class $C^{(1)}$. In [18], (1.3) is considered with $\int_a^\infty |dv_1|<\infty$, $\int_a^\infty |v_3|<\infty$, and $v_2\equiv 0$. In [17], (1.1) is considered with $p=w\equiv 1$ and q sufficiently smooth.

Also related to the resutls here are those of M. Ben-Artzi and A. Devinatz [2] and M. Ben-Artzi [3], [4]. In [2], [3], [4], the operator $-\Delta+V$ is considered on R^n where V is spherically symmetric. In addition to the question of absolute continuity of the spectrum, the existence and completeness of wave operators is discussed. The ordinary differential operators considered are defined on $(0,\infty)$, with both endpoints being singular. In [2], [3] an operator of type (1.3) is considered. The singularity at infinity has $v_2=0$, but allows a v_1 term more general than that considered here. The singularity at zero has more stringent growth conditions than that required by Theorem 4.2 below. In [4], similar questions are discussed for (1.1) with $p=w=1$ and q sufficiently smooth.

Additional criteria for the continuous spectrum to be absolutely continuous may be found in P. A. Rejto [13], [15], and P. A. Rejto and K. Sinha [14]. These operators are of type (1.1) with $p=w=1$, but with two singular endpoints. In [13], [15] operators are considered on $(0,\infty)$, and in [14] they are considered on $(-\infty,\infty)$. The method in these papers is to apply an abstract criterion for absolute continuity based on the resolvent operator. In [2], [3], [4] the criterion is similar but based on the Green's function. The method employed here uses a numerical component of the Green's function—the Titchmarsh–Weyl coefficient whose singular structure yields the various components of the spectrum (cf. [6]).

For two singular endpoint problems we follow the method of [10] which treats each singular endpoint separately. This allows greater flexibility in placing constraints on the coefficients. This permits, for example, a generalization of [13, Thm. 2.1] along the lines of §6 where the singularity at zero is transformed to infinity.

A problem related to this paper is showing the absence of eigenvalues embedded in the continuous spectrum. Some sharp bounds for cutoff of eigenvalues may be found in the paper of F. V. Atkinson and W. N. Everitt [1].

In §2 below we define a generalization of (1.3) and show how it relates to transformed equations. In §3 we develop the necessary asymptotics for solutions. In §4, the Titchmarsh–Weyl m-coefficient is employed to obtain absolute continuity of the continuous spectrum. In §5, the theory of §4 is applied to analyze the two-singular endpoint problem, and a connection is made with recent work of R. Carmona [5]. In §6, the theory is applied to the hydrogen atom.

By the norm of a vector we mean the Euclidean norm; by the norm of a matrix we mean the corresponding operator norm. Standard notation is used for the components of vectors and matrices. For $w(x)>0$ and Lebesgue measurable, $\mathscr{L}_w^p(I)$ denotes the Banach space of all equivalence classes of complex-valued functions f satisfying $\int_I w|f|^p < \infty$.

2. A general equation. The equation considered is

$$y''+\left[a(\lambda)^2+v_1(x,\lambda)+v_2(x,\lambda)+v_3(x,\lambda)\right]y=0, \qquad a\leqq x<\infty, \tag{2.1}$$

where the conditions on a, v_1, v_2, and v_3 are given below. To motivate the form of (2.1) considered, we return to (1.1) with a singular point at b, $b\leqq\infty$, i.e.

$$-(py')'+qy=\lambda wy, \qquad a\leqq x<b. \tag{2.2}$$

Suppose in (2.2) we make the Kummer–Liouville change of variables,

$$y(x)=h(x)z(t), \quad h=(pw)^{-1/4}, \quad t=f(x)=\int_a^x\left(\frac{w}{p}\right)^{1/2},$$

where p,w are positive and sufficiently smooth and $\int_a^b (w/p)^{1/2}=\infty$. Then (2.2) becomes

$$-\ddot{z}(t)+Q(t)z(t)=\lambda z(t), \qquad 0\leqq t<\infty, \tag{2.3}$$

with $Q(t)=[q/w-h(ph')'/f'](x)$. For appropriate p,q, and w, (2.3) is of type (2.1). Using the Titchmarsh–Weyl m-coefficient below to analyze the spectrum of (2.2), we need only know the asymptotics of (2.3).

Suppose now in (2.2), $q=-n-r$ where $n(x)>0$ is a smooth part of q and r is a perturbation part. If we make the change of variables

$$y(x)=h(x)z(t), \quad h=(pn)^{-1/4}, \quad t=f(x)=\int_a^x\left(\frac{n}{p}\right)^{1/2}$$

where $\int_a^\infty (n/p)^{1/2}=\infty$, then (2.2) becomes

$$-\ddot{z}(t)-[1+Q(t)]z(t)=0, \qquad 0\leqq t<\infty, \tag{2.4}$$

with $Q(t)=[r/n+h(ph')'/f'+\lambda w/n](x)$; hence the λ-dependence is changed to a v term in (2.1), and $a(\lambda)$ is constant. The case $n(x)<0$ above is not considered since it typically leads to an empty continuous spectrum.

Let $C_+ = \{\lambda | \operatorname{Im}\lambda \geqq 0\}$ and $C_0 = \{\lambda \in C_+ | \operatorname{Re}\lambda \geqq 0\}$. Our assumptions for (2.1) are:

(A_1) For $i=1,2,3$, $v_i(x,\lambda) = v_{i1}(x) + \lambda v_{i2}(x)$ where each v_{ij} is real and locally Lebesgue integrable. Further

a) For $j=1,2$, $v_{1j}(x) \to 0$ as $x \to \infty$, v_{1j} is absolutely continuous with $v'_{1j} \in \mathscr{L}(a,\infty)$; $v_{12}(x) \geqq 0$.

b) For $j=1,2$, $\int_a^\infty v_{2j}$ exists (conditionally), $V_{2j}(x) = \int_x^\infty v_{2j}$ is in $\mathscr{L}(a,\infty)$; $W_{sj}(x) = \int_x^\infty |v'_{1s} V_{2j}|$ is in $\mathscr{L}(a,\infty)$ for $s,j=1,2$.

(c) For $j=1,2$, $v_{3j} \in \mathscr{L}(a,\infty)$.

(A_2) $a(\lambda)$ is continuous on C_+, analytic on the interior of C_+, and $a(C_+) \subset C_0$. There is a number $\lambda_0 \in C_+$ such that if U is a closed subset of C_+ not containing λ_0, then $a(U)$ is bounded away from zero.

Note that if U is as in (A_2), then $|a(\lambda)|^2 \geqq \varepsilon > 0$ on U for some $\varepsilon > 0$. Now if U is compact or $v_{12}(x) \equiv 0$ we may, by redefining v_3 if necessary, assume $|v_1(x,\lambda)| \leqq \varepsilon/2$ for $a \leqq x < \infty$, $\lambda \in U$. Without loss of generality we will always make this assumption for U in (A_2) compact or $v_{12}(x) \equiv 0$.

3. Asymptotic theory of (2.1). We suppose in this section that (A_1)–(A_2) hold and U is a closed subset of C_+ not containing λ_0, and that either U is compact or $v_{12}(x) \equiv 0$. Hence

$$K = K(x,\lambda) = \left[a(\lambda)^2 + v_1(x,\lambda)\right]^{1/2} \tag{3.1}$$

is a well-defined element of C_0 and satisfies for some $\varepsilon > 0$

$$|K(x,\lambda)| \geqq \left(\frac{\varepsilon}{2}\right)^{1/2}, \qquad \lambda \in U, \quad a \leqq x < \infty. \tag{3.2}$$

Write (2.1) in system form as

$$\begin{pmatrix} y \\ y' \end{pmatrix}' = \left\{ \begin{pmatrix} 0 & 1 \\ -K^2 & 0 \end{pmatrix} - (v_2 + v_3) \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} \right\} \begin{pmatrix} y \\ y' \end{pmatrix}. \tag{3.3}$$

First we transform (3.3) as in [10]. Set

$$S = \begin{pmatrix} 1 & 1 \\ iK & -iK \end{pmatrix}, \qquad \psi = S^{-1} \begin{pmatrix} y \\ y' \end{pmatrix} = \frac{1}{2} \begin{pmatrix} 1 & 1/iK \\ 1 & -1/iK \end{pmatrix} \begin{pmatrix} y \\ y' \end{pmatrix};$$

hence (3.3) becomes

$$\psi' = \left\{ iK \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} - \left(\frac{v_2 + v_3}{2iK}\right) \begin{pmatrix} 1 & 1 \\ -1 & -1 \end{pmatrix} - \frac{K'}{2K} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \right\} \psi. \tag{3.4}$$

Define

$$N = \begin{pmatrix} 1 & 1 \\ -1 & -1 \end{pmatrix}, \qquad R = \left(\frac{-v_3}{2iK}\right) N - \frac{K'}{2K} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix},$$

$$E = E(x,\lambda) = \begin{pmatrix} \exp\left(\int_a^x iK\right) & 0 \\ 0 & \exp\left(\int_a^x -iK\right) \end{pmatrix},$$

$$W = W(x,\lambda) = \int_x^\infty \frac{v_2}{2iK}.$$

Then by the variation of constants formula, (3.4) can be written as

$$\psi(x,\lambda)=E(x,\lambda)\Big\{\psi(a)+\int_a^x E^{-1}[W'N+R]\psi\Big\}. \tag{3.5}$$

Integrating the term $\int_a^x E^{-1}W'N\psi$ in (3.5) by parts and using $\psi'=[E'E^{-1}+W'N+R]\psi$, we obtain after simplifying that

$$\begin{aligned}&[I-W(x,\lambda)N]\psi(x,\lambda)\\ &\quad=E(x,\lambda)\Big\{[I-W(a,\lambda)N]\psi(a,\lambda)\\ &\qquad+\int_a^x E^{-1}[R+E'E^{-1}WN-WNE'E^{-1}]\psi\Big\}.\end{aligned} \tag{3.6}$$

Equation (3.6) simplifies to

$$\begin{aligned}&[I-W(x,\lambda)N]\psi(x,\lambda)\\ &\quad=E(x,\lambda)\Big\{[I-W(a,\lambda)N]\psi(a,\lambda)+\int_a^x E^{-1}R_1\psi\Big\}\end{aligned} \tag{3.7}$$

where

$$R_1=R_1(x,\lambda)=R(x,\lambda)+2iK(x,\lambda)W(x,\lambda)\begin{pmatrix}0&1\\1&0\end{pmatrix}. \tag{3.8}$$

The definition of W and an integration by parts yields that

$$W(x,\lambda)=\frac{V_2(x,\lambda)}{2iK(x,\lambda)}-\int_x^\infty (V_2v_1'/4iK^3) \tag{3.9}$$

where $V_2(x,\lambda)=\int_x^\infty v_2$. Thus by (A_1)–(A_2) we have for each compact set F in C_+ a constant M_F such that for all $\lambda\in U\cap F$,

$$\int_a^\infty \|R_1(x,\lambda)\|\,dx\leqq M_F,\qquad |W(x,\lambda)|\leqq M_F. \tag{3.10}$$

In case $v_{2j}(x)\equiv 0$ for $j=1,2,3$, (3.10) holds for $F=C_+$. Set

$$\begin{pmatrix}c_1\\c_2\end{pmatrix}=[I-W(a,\lambda)N]\psi(a,\lambda),\qquad \psi_1(x,\lambda)=\Big[\exp\Big(\int_a^x iK\Big)\Big]\psi(x,\lambda). \tag{3.11}$$

Then (3.7) becomes

$$\begin{aligned}&[I-W(x,\lambda)N]\psi_1(x,\lambda)\\ &\quad=\begin{pmatrix}\exp\Big(2i\int_a^x K\Big)&0\\0&1\end{pmatrix}\begin{pmatrix}c_1\\c_2\end{pmatrix}\\ &\qquad+\int_a^x\begin{pmatrix}\exp\Big(2i\int_s^x K\Big)&0\\0&1\end{pmatrix}R_1(s,\lambda)\psi_1(x,\lambda)\,ds.\end{aligned} \tag{3.12}$$

Note that by (3.10), for $\lambda\in U\cap F$.

$$\big\|[I-W(x,\lambda)N]^{-1}\big\|=\|[I+W(x,\lambda)N]\|\leqq 1+M_F\|N\|. \tag{3.13}$$

LEMMA 3.1. *Suppose* (A_1)–(A_2) *hold,* U *is as above,* $F \subset C_+$ *is compact and* (3.11) *holds. Then for* $\lambda \in U \cap F$ *and* $a \leqq x < \infty$,

$$\|\psi_1(x,\lambda)\| \leqq \gamma \left\| \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} \right\| e^{\gamma M_F} \tag{3.14}$$

where $\gamma = 1 + M_F \|N\|$.

Proof. Recall $\operatorname{Im} K(x,\lambda) \geqq 0$. Thus by (3.12) and (3.13),

$$\|\psi_1(x,\lambda)\| \leqq \gamma \left[\left\| \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} \right\| + \int_a^x \|R_1(x,\lambda)\| \|\psi_1(s,\lambda)\| \, ds \right].$$

Equation (3.14) now follows by an application of Gronwall's inequality.

Lemma 3.1 yields asymptotic behavior of ψ_1 as $x \to \infty$ and hence of y, y'. First we rewrite (3.12) as

(3.15)

$$[I - W(x,\lambda)N]\psi_1(x,\lambda) = \begin{pmatrix} 0 \\ A(\lambda) \end{pmatrix} - \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \int_x^\infty R_1\psi_1 + \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \left\{ \left[\exp\left(2i \int_a^x K \right) \right] \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} + \int_a^x \left[\exp\left(2i \int_s^x K \right) \right] R_1\psi_1 \right\}$$

where

$$\begin{pmatrix} 0 \\ A(\lambda) \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \left\{ \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} + \int_a^\infty R_1\psi_1 \right\}. \tag{3.16}$$

LEMMA 3.2. *Suppose* (A_1)–(A_2) *hold,* U *is as above, and* $y(\cdot,\lambda)$ *is a solution of* (2.1) *with* $y(0,\lambda)$ *and* $y'(0,\lambda)$ *independent of* λ. *Then* $A(\lambda) = A_y(\lambda)$ *defined by* (3.16) *is continuous in* λ *for* λ *in* U *and analytic on the interior of* U.

Proof. The method of successive approximations for existence of solutions shows $y(x,\lambda)$ is jointly continuous in x and λ and entire in λ for fixed x. Since K is analytic in λ for λ in the interior of U and continuous in λ for λ in U, the functions $R_1(x,\lambda)$ and $\psi_1(x,\lambda)$ have the same properties for fixed x; further they are jointly continuous in x and λ. Hence by the bounds (3.10) and (3.14), $A(\lambda)$ is the limit of a sequence of functions which are (i) continuous on U, (ii) analytic on interior U, and (iii) uniformly bounded on compact sets. By Vitali's theorem, $A(\lambda)$ is analytic on interior U. Further, (A_1) implies that $\int_n^\infty \|R_1(s,\lambda)\| ds \to 0$ as $n \to \infty$, uniformly for λ in compact sets. Thus $A(\lambda)$ is continuous on U.

Fix now $\lambda \in U$. If $a(\lambda)$ is in the interior of C_0 or $a(\lambda)$ is on the positive imaginary axis, then $K(x,\lambda) \to a(\lambda)$ as $x \to \infty$. Then $\exp \int_a^x iK \to 0$ as $x \to \infty$, and we have the following asymptotic behavior. (Note that $W(x,\lambda) \to 0$ as $x \to \infty$.) As $x \to \infty$,

$$\begin{aligned} \psi_1(x,\lambda) &= \begin{pmatrix} 0 \\ A(\lambda) \end{pmatrix} + o(1), \\ \begin{pmatrix} y \\ y' \end{pmatrix}(x,\lambda) &= \left[\exp\left(- \int_a^x iK \right) \right] A(\lambda) \begin{pmatrix} 1 + o(1) \\ -iK + o(1) \end{pmatrix}. \end{aligned} \tag{3.17}$$

If $a(\lambda)$ is in the positive real axis, then

$$K = \left(a(\lambda)^2 + v_1 \right)^{1/2} = a(\lambda)\left(1 + v_1 / a(\lambda)^2 \right)^{1/2}$$

from which we conclude that

$$\int_a^\infty \operatorname{Im} K(x,\lambda)\,dx=\infty \Leftrightarrow \int_a^\infty \operatorname{Im} v_1(x,\lambda)\,dx=\infty.$$

If $\int_a^\infty \operatorname{Im} v_1=\infty$, then (3.17) holds. If $\int_a^\infty \operatorname{Im} v_1<\infty$, then from (3.15) we conclude that as $x\to\infty$,

$$\psi_1(x,\lambda)=\begin{pmatrix}0\\ A(\lambda)\end{pmatrix}+\exp\left(2i\int_a^x K\right)\begin{pmatrix}B(\lambda)\\ 0\end{pmatrix}+o(1), \tag{3.18}$$

$$\begin{pmatrix}y\\ y'\end{pmatrix}(x,\lambda)=\exp\left(-\int_a^x iK\right)A(\lambda)\begin{pmatrix}1+o(1)\\ -iK+o(1)\end{pmatrix}$$

$$+\exp\left(\int_a^x iK\right)B(\lambda)\begin{pmatrix}1+o(1)\\ iK+i(1)\end{pmatrix},$$

where

$$\begin{pmatrix}B(\lambda)\\ 0\end{pmatrix}=\begin{pmatrix}1&0\\ 0&0\end{pmatrix}\left\{\begin{pmatrix}c_1\\ c_2\end{pmatrix}+\int_0^\infty \exp\left(-2i\int_a^s K\right)R_1\psi_1\right\}.$$

If now in (3.18), $y(x,\lambda)$ is real and $K(x,\lambda)$ is real, then from $\operatorname{Im} y(x,\lambda)\equiv 0$ we conclude that $B(\lambda)=\overline{A(\lambda)}$.

4. Application to half-line problems. Consider first equation (1.1) with the boundary condition (1.2). Define a fundamental set of solutions θ_α, ϕ_α of (1.1) by the initial values

$$\begin{pmatrix}\theta_\alpha & \phi_\alpha\\ p\theta_\alpha' & p\phi_\alpha'\end{pmatrix}(a,\lambda)=\begin{pmatrix}\sin\alpha & -\cos\alpha\\ \cos\alpha & \sin\alpha\end{pmatrix}. \tag{4.1}$$

If L is in the limit point case at infinity, then the limit, for $\operatorname{Im}\lambda\neq 0$,

$$\lim_{x\to\infty} -\frac{\theta_\alpha(x,\lambda)}{\phi_\alpha(x,\lambda)}=m_\alpha(\lambda) \tag{4.2}$$

exists and is analytic on $\operatorname{Im}\lambda\neq 0$. Further the self-adjoint operator T determined by (1.1)–(1.2) has spectral function ρ_α related to m_α by

$$\rho_\alpha(t)-\rho_\alpha(s)=\lim_{\varepsilon\downarrow 0}\frac{1}{\pi}\int_s^t \operatorname{Im} m_\alpha(u+i\varepsilon)\,du \tag{4.3}$$

at points of continuity s,t of ρ_α. (cf. [7, Chap. 9]).

Detailed relations between the singular structure of m_α and the spectrum have been established by J. Chandhuri and W. N. Everitt [6]. In particular we note that the isolated eigenvalues coincide with the poles of m_α.

In this section we solve the half-line problem for (1.3) and then show how the Kummer–Liouville transformation, together with the asymptotic theory of §3 solves a wider class of problems. For (1.3), we take in (A_1)–(A_2) for some real λ_0 that

$$a(\lambda)=\sqrt{\lambda-\lambda_0}, \qquad v_{12}=v_{22}=v_{32}\equiv 0. \tag{4.4}$$

The set U used in §3 may be taken to be any closed subset of C_+ not containing λ_0. Let A_ϕ, A_θ be the functions defined by (3.16) for ϕ_α, θ_α respectively. Then A_ϕ, A_θ are continuous on $C_+\backslash\{\lambda_0\}$ and analytic on the interior of this set. Further for λ real, $\lambda<\lambda_0$, iK and hence A_ϕ, A_θ are real. Thus by reflection, A_ϕ, A_θ have analytic continuations into $\{\lambda\colon \mathrm{Re}\lambda<\lambda_0\}$.

For real λ, $\lambda>\lambda_0$, write

$$A_\phi = a_\phi + ib_\phi, \qquad A_\theta = a_\theta + ib_\theta, \tag{4.5}$$

where a_ϕ, b_ϕ, a_θ, b_θ are real. From the asymptotic formula (3.18) we have as $x\to\infty$, $\lambda>\lambda_0$,

$$\begin{aligned}\phi_\alpha(x,\lambda) &= 2a_\phi(\lambda)\cos\int_a^x K + 2b_\phi(\lambda)\sin\int_a^x K + o(1),\\ \phi'_\alpha(x,\lambda) &= 2K(x,\lambda)\Big[-a_\phi(\lambda)\sin\int_a^x K + 2b_\phi(\lambda)\cos\int_a^x K\Big] + o(1),\end{aligned} \tag{4.6}$$

with similar formulas for θ_α, θ'_α. From the Wronskian,

$$1\equiv\begin{vmatrix}\theta_\alpha & \phi_\alpha\\ \theta'_\alpha & \phi'_\alpha\end{vmatrix}(x,\lambda) = 4K(x,\lambda)\big[a_\theta(\lambda)b_\phi(\lambda)-a_\phi(\lambda)b_\theta(\lambda)\big]+o(1),$$

we conclude that for $\lambda>\lambda_0$,

$$a_\theta(\lambda)b_\phi(\lambda)-a_\phi(\lambda)b_\theta(\lambda) = \frac{1}{4(\lambda-\lambda_0)^{1/2}}. \tag{4.7}$$

This shows that $A_\phi(\lambda)$, $A_\theta(\lambda)$ do not vanish for real $\lambda>\lambda_0$ and hence also for λ sufficiently near this set. For $\mathrm{Im}\lambda>0$ such that $A_\phi(\lambda)\neq 0$, we have from the asymptotic formula (3.18) and (4.2) that

$$m_\alpha(\lambda) = -\lim_{x\to\infty}\frac{\exp(-i\int_a^x K)\big[A_\theta(\lambda)+o(1)\big]}{\exp(-i\int_a^x K)\big[A_\phi(\lambda)+o(1)\big]} = -\frac{A_\theta(\lambda)}{A_\phi(\lambda)}. \tag{4.8}$$

Recall that A_θ, A_ϕ are analytic on $\mathrm{Im}\lambda>0$ and $\mathrm{Re}\lambda<\lambda_0$ and that they are continuous on $C_+\backslash\{\lambda_0\}$. From (4.5), (4.7), and (4.8), we have that for $\mu>\lambda_0$,

$$\begin{aligned}\lim_{\varepsilon\downarrow 0+}\mathrm{Im}\,m_\alpha(\mu+i\varepsilon) &= -\mathrm{Im}\,\frac{A_\theta(\mu)}{A_\phi(\mu)}\\ &= -\frac{-a_\theta b_\phi + a_\phi b_\theta}{a_\phi^2+b_\phi^2}(\mu)\\ &= \frac{1}{4(\mu-\lambda_0)^{1/2}\big[a_\phi^2(\mu)+b_\phi^2(\mu)\big]}.\end{aligned} \tag{4.9}$$

Since $A_\phi(\mu)\neq 0$, and the limit in (4.9) is uniform on compact subintervals of (λ_0,∞), we have from (4.3) that for $\mu>\lambda_0$, ρ_α is class $C^{(1)}$ with

$$\rho'_\alpha(\mu) = \frac{1}{4(\mu-\lambda_0)^{1/2}\big[a_\phi^2(\mu)+b_\phi^2(\mu)\big]}.$$

On the other hand A_θ, A_ϕ are real for $\mu<\lambda_0$. Thus $m_\alpha(\lambda)$ is meromorphic on $\mathrm{Re}\lambda<\lambda_0$, and on a compact subinterval of $(-\infty,\lambda_0)$, ρ_α is a step function with a finite number of jumps.

THEOREM 4.1. *Assume in* (1.3) *that* v_1, v_2, v_3 *are locally Lebesgue integrable and* (i) $v_1(x)\to-\lambda_0$ *as* $x\to\infty$. (ii) $V_2(x)=\int_x^\infty v_2$ *exists, and* (iii) v_1', v_3, V_2, *and* W_0 *are in* $\mathscr{L}(a,\infty)$ *where* $W_0(x)=\int_x^\infty |v_1'V_2|$. *Let* $0\leqq\alpha<\pi$ *and* T_α *in* $\mathscr{L}^2(a,\infty)$ *be defined by*

$$T_\alpha(y)=-(y''+[v_1+v_2+v_3]y)$$

with domain T_α *being all* $y\in\mathscr{L}^2(a,\infty)$ *such that* y,y' *are locally absolutely continuous,* $T_\alpha(y)\in\mathscr{L}^2(a,\infty)$, *and* y *satisfies the boundary condition* $\sin\alpha\, y(a)+\cos\alpha\, y'(a)=0$. *Then* T_α *is self-adjoint, on* (λ_0,∞) *its spectral function* ρ_α *is* $C^{(1)}$ *with* $\rho_\alpha'(\mu)>0$, *on each compact subinterval of* $(-\infty,\lambda_0)$, *the spectrum of* T_α *is either empty or consists of a finite number of eigenvalues, and the spectrum of* T_α *is bounded below.*

Proof. The asymptotic form of solutions (4.6) shows that for $\lambda>\lambda_0$, (1.3) has no solutions in $\mathscr{L}^2(a,\infty)$. Thus T_α is of limit point type at infinity, and the theory of self-adjoint extensions of symmetric operators shows T_α is self-adjoint. The only part of Theorem 4.1 not proved above is that the spectrum is bounded below. We now prove that under the above assumptions, a solution y of (1.3), with $y(a\lambda)$ and $y'(a,\lambda)$ independent of λ, satisfies

$$\lim_{\lambda\to-\infty} A_y(\lambda)=y(a)/2 \tag{4.10}$$

where $A(\lambda)=A_y(\lambda)$ is given by (3.16). First we show that (4.10) implies the spectrum of T_α is bounded below. Since m_α is meromorphic on $\mathrm{Re}\lambda<\lambda_0$, the conclusion follows by showing m_α has a limit as $\lambda\to-\infty$. For $\alpha\neq\pi/2$, (4.10) implies $m_\alpha(\lambda)\to -\theta_\alpha(a,\lambda)/\phi_\alpha(a,\lambda)=\tan\alpha$ as $\lambda\to-\infty$. For $\alpha=\pi/2$, $\phi_\alpha(a,\lambda)=0$ and $\theta_\alpha(a,\lambda)=1$ so that (4.10) implies $m_\alpha(\lambda)^{-1}\to0$ as $\lambda\to-\infty$.

Thus the zeros of $m_\alpha(\lambda)$ are bounded below; hence the poles of $m_\alpha(\lambda)$ are bounded below since $m_\alpha(\lambda)$ is real on $(-\infty,\lambda_0)$. Further, $\mathrm{Im}\, m_\alpha(\lambda)>0$ for $\mathrm{Im}\lambda>0$, and we have by orientation preserving properties of analytic functions that $m_\alpha(\lambda)\to -\infty$ as $\lambda\to-\infty$.

To establish (4.10) we first note that in this case the bounds (3.10) hold for $F=C_+$, $U=\{\lambda\in C_+\,|\,\mathrm{Re}\lambda\leqq\lambda_0-1\}$. This is because of the way λ enters the various terms of $R_1(x,\lambda)$.

From (3.11) and (3.16)

$$A_y(\lambda)=\left[\frac{1}{2}+W(a,\lambda)\right]y(a)-\frac{1}{2}\frac{y'(a)}{iK(a,\lambda)}+\int_a^\infty (R_1\psi_1)_2. \tag{4.11}$$

The definition of K and (3.9) show that for each x,

$$\lim_{\lambda\to-\infty} W(x,\lambda)=0.$$

The definition of R shows $\int_a^\infty \|R(x,\lambda)\|dx\to0$ as $\lambda\to-\infty$. Since ψ_1 is uniformly bounded on $a\leqq x<\infty$, $\lambda\leqq\lambda_0-1$, we conclude that

$$\int_a^\infty \|R(x,\lambda)\psi_1(x,\lambda)\|\,dx\to0 \quad \text{as } \lambda\to\infty.$$

Thus to complete the proof of (4.10), we have from (4.11) and (3.8) that it is sufficient to prove as $\lambda\to-\infty$,

$$\int_a^\infty \left\{2iKW\begin{pmatrix}0&1\\1&0\end{pmatrix}\psi_1\right\}_2=\int_a^\infty 2iKW(\psi_1)_1\to0. \tag{4.12}$$

Since $iK(x,\lambda)=i(\lambda-\lambda_0+v_1(x))^{1/2}$, we have from the equation (3.15) that $(\psi_1(x,\lambda))_1\to 0$ as $\lambda\to-\infty$. Now from (3.9), we see that there is a constant k such that for $a\leqq x<\infty$, $-\infty<\lambda\leqq\lambda_0-1$,

$$|K(x,\lambda)W(x,\lambda)|\leqq k\left[|V_2(x)|+W_0(x)\right]. \tag{4.13}$$

Since $(\psi_1(x,\lambda))_1$ is uniformly bounded for $a\leqq x<\infty$, $-\infty<\lambda\leqq\lambda_0-1$, and for fixed x tends to zero as $\lambda\to-\infty$, application of (4.13) and the Lebesgue dominated convergence theorem yields (4.12).

An example of an equation of type (2.2) to which Theorem 4.1 applies is

$$-y''+\left[\frac{c}{x^2(\ln x)}+r(x)\right]y=\lambda x^{-2}y,\qquad e\leqq x<\infty, \tag{4.14}$$

where x is a constant and $\int_e^\infty x|r(x)|dx<\infty$. The transformed equation (2.3) is ($t=\ln x, h(x)=x^{1/2}, y(x)=h(x)z(t)$),

$$-\ddot{z}(t)+\left[\frac{c}{t}+x^2r(x)+\frac{1}{4}\right]z(t)=\lambda z(t),\qquad 1\leqq t<\infty,$$

which satisfies the conditions of Theorem 4.1. A calculation shows that the change of variable is a unitary map from $\mathscr{L}^2_w(e,\infty)$ onto $\mathscr{L}^2(1,\infty)$. Thus the spectral function ρ of a self-adjoint operator associated with (4.14) satisfies $\rho'(\mu)>0$ on $(\frac{1}{4},\infty)$.

Examples given by M. S. P. Eastham and H. Kalf [8, p. 89] show that v terms in (1.3) cannot be much larger without possibly introducing an eigenvalue in the continuous spectrum. As shown in [8], the equation, on $1\leqq x<\infty$,

$$y''(x)+\left[1-4x^{-1}\sin 2x+x^{-2}(2\cos^2 x+4\cos^4 x)\right]y=0$$

has a solution $y(x)=O(x^{-1})$. Thus $\lambda=1$ is an eigenvalue in the continuous spectrum $[0,\infty)$ for an appropriate boundary condition imposed at $x=1$.

As a second application of the asymptotic theory of §3, we allow the potential to go to infinity.

THEOREM 4.2. *Suppose p,n,r, and w are real locally Lebesgue integrable functions on $[a,b)$ with p,n, and w positive. Suppose p and n are of class $C^{(2)}[a,\infty)$, $\int_a^b (n/p)^{1/2}=\infty$, and $\int_a^b wh^2=\infty$ where $h=(pn)^{-1/4}$. Set $t=f(x)=\int_a^x (n/p)^{1/2}$ and suppose*

$$\left[\frac{h(ph')'}{f'}+\frac{r}{n}\right](x)=v_{11}(t)+v_{21}(t)+v_{31}(t),$$
$$\left(\frac{w}{n}\right)(x)=v_{12}(t)+v_{22}(t)+v_{32}(t)$$

where the $v_{ij}(t)$ satisfy the hypothesis (A_1). *Define an operator T_α in $\mathscr{L}^2_w(a,b)$ by $T_\alpha(y)=w^{-1}[-(py')'-(n+r)y]$ where the domain of T_α is the set of all y in $\mathscr{L}^2_w(a,b)$ such that y,y' are locally absolutely continuous, $T_\alpha(y)\in\mathscr{L}^2_w(a,b)$, and y satisfies the boundary condition $\sin\alpha\, y(a)+\cos\alpha\,(py')(a)=0$. Then T_α is self-adjoint and its spectral function ρ_α is of class $C^{(1)}(-\infty,\infty)$ with $\rho'_\alpha(\mu)>0$ for all real μ.*

Proof. As in (2.2), make the transformation

$$y(x)=h(x)z(t),\quad h=(pn)^{-1/4},\quad t=f(x)=\int_a^x (n/p)^{1/2}, \tag{4.15}$$

so that z satisfies

$$\ddot{z}(t)+[1+Q(t)]z(t)=0 \tag{4.16}$$

with, using the above notation,

$$Q(t)=[v_{11}(t)+v_{21}(t)+v_{31}(t)]+\lambda[v_{12}(t)+v_{22}(t)+v_{32}(t)].$$

With $a(\lambda)=1$, the theory of §3 is applicable. Since

$$\int_0^\infty v_{12}(t)\,dt=\int_a^b (w/n)(n/p)^{1/2}\,dx-\int_0^\infty [v_{22}(t)+v_{32}(t)]\,dt=\infty,$$

we obtain from (3.17) and (3.18) that for a solution $z(t,\lambda)$ of (4.16), as $t\to\infty$,

$$\begin{aligned}
&\operatorname{Im}\lambda>0\colon z(t,\lambda)=\exp\Big(-\int_0^t iK\Big)[A_z(\lambda)+o(1)],\\
&\operatorname{Im}\lambda=0\colon z(t,\lambda)=\exp\Big(-\int_0^t iK\Big)[A_z(\lambda)+o(1)],\\
&\qquad+\exp\Big(\int_0^t iK\Big)[\overline{A_z(\lambda)}+o(1)].
\end{aligned}\tag{4.17}$$

As in the proof of Theorem 4.1, it follows that $A_z(\lambda)\neq 0$ for λ real. Since the Wronskian of two solutions of (4.16) is constant, it follows from the asymptotic form of $z,\dot z$ for $\operatorname{Im}\lambda=0$ that $A_{z_1}(\lambda)/A_{z_2}(\lambda)$ is not real for linearly independent solutions z_1, z_2 of (4.16). After some calculations it follows that for some $\delta>0$, $0\leqq t<\infty$,

$$|z_1(t,\lambda)|^2+|z_2(t,\lambda)|^2\geqq\delta.$$

If y_1 and y_2 are the corresponding solutions of (2.2),

$$\int_a^b w\left[|y_1|^2+|y_2|^2\right]\geqq\delta\int_a^b w(pn)^{-1/2}=\infty.$$

Thus (2.2) is of limit point type at b; hence T_α as defined above is self-adjoint.

Consider the solutions θ_α, ϕ_α of (2.2) defined by the initial conditions

$$\begin{pmatrix}\theta_\alpha & \phi_\alpha\\ \theta_\alpha' & \phi_\alpha'\end{pmatrix}(a,\lambda)=\begin{pmatrix}\sin\alpha & -\cos\alpha\\ \cos\alpha & \sin\alpha\end{pmatrix}$$

and let z_θ, z_ϕ be the corresponding solutions of (4.16). Now for $\operatorname{Im}\lambda>0$ in (4.17),

$$m_\alpha(\lambda)=-\lim_{x\to b}\frac{\theta_\alpha(x,\lambda)}{\phi_\alpha(x,\lambda)}=-\lim_{t\to\infty}\frac{z_\theta(t,\lambda)}{z_\phi(t,\lambda)}=\frac{-A_{z_\theta}(\lambda)}{A_{z\phi}(\lambda)}.\tag{4.18}$$

Since A_{z_θ}, A_{z_ϕ} are continuous on C_+ and do not vanish for λ real, it follows as in (4.9) that $\operatorname{Im} m_\alpha(\mu+i\varepsilon)$ has a positive limit as $\varepsilon\downarrow 0$. This completes the proof.

The proof shows that the condition $\int_a^b w(pn)^{-1/2}=\infty$ is required to prevent the equation (2.2) from being of limit circle type at b. In this case the essential spectrum would be empty. It is also required to produce the asymptotic behavior (4.17) for $\operatorname{Im}\lambda>0$ which in turn implies (4.18).

An example of an equation which satisfies the hypothesis of Theorem 4.2 is:

$$y''+[\lambda+x^2+ax^\alpha\sin x^\beta+bxf(x)]y=0,\qquad 1\leqq x<\infty,$$

with $\beta>\alpha+2$, $\int_1^\infty|f|<\infty$, a and b constants.

5. Application to whole-line problems. We consider here only the equation

$$y''+[\lambda+v_1(x)+v_2(x)+v_3(x)]y=0, \qquad -\infty<x<\infty, \tag{5.1}$$

where the v_i satisfy the conditions of Theorems 4.1 at each of $\pm\infty$. The methods employed here will apply to other classes of equations. Following [10], let

$$v_{\pm}=\lim_{x\to\pm\infty} -v_1(x).$$

Since (5.1) is in the limit point case at $\pm\infty$, a self-adjoint operator T is defined in $\mathscr{L}^2(-\infty,\infty)$ by

$$T(y)=-(y''+[v_1(x)+v_2(x)+v_3(x)]y)$$

where the domain of T consists of all $y\in\mathscr{L}^2(-\infty,\infty)$ with y,y' locally absolutely continuous and $T(y)\in\mathscr{L}^2(-\infty,\infty)$. We define a basis θ,ϕ of (5.1) by the initial values

$$\begin{pmatrix}\theta & \phi\\ \theta' & \phi'\end{pmatrix}(0,\lambda)=\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}.$$

The Titchmarsh–Weyl m-coefficients $m_{\pm}$ are defined on $\operatorname{Im}\lambda>0$ by

$$m_{\pm}(\lambda)=-\lim_{x\to\pm\infty}\frac{\theta(x,\lambda)}{\phi(x,\lambda)}.$$

Then for $\operatorname{Im}\lambda>0$, $\operatorname{Im}m_+(\lambda)>0$ and $\operatorname{Im}m_-(\lambda)<0$ (cf. [9]). The spectral function ρ of T is a 2×2 matrix and depends on the choice of basis [7, Chap. 9]. However, any two such ρ's are similar [9]. For the basis above, we have at points of continuity t,s,

$$\rho(t)-\rho(s)=\lim_{\varepsilon\downarrow0}\frac{1}{\pi}\int_s^t \operatorname{Im}M(u+i\varepsilon)\,du \tag{5.2}$$

where M is given by

$$M=\begin{pmatrix}1/(m_--m_+) & (m_++m_-)/2(m_--m_+)\\ (m_++m_-)/2(m_--m_+) & m_+m_-/(m_--m_+)\end{pmatrix} \tag{5.3}$$

$$=\begin{pmatrix}m_{11} & m_{12}\\ m_{21} & m_{22}\end{pmatrix}.$$

Suppose we write for $\operatorname{Im}\lambda>0$, with a,b,c,d real,

$$m_+(\lambda)=a(\lambda)+ib(\lambda), \qquad m_-(\lambda)=c(\lambda)+id(\lambda).$$

Then a straightforward calculation yields

$$\begin{aligned}
&\operatorname{Im}m_{11}=\frac{b-d}{D}, \qquad D=(c-a)^2+(b-d)^2,\\
&\operatorname{Im}m_{12}=\operatorname{Im}m_{21}=\frac{bc-ad}{D},\\
&\operatorname{Im}m_{22}=\frac{bc^2-a^2d+bd^2-b^2d}{D},\\
&\det\operatorname{Im}M=\frac{-bd}{D}.
\end{aligned} \tag{5.4}$$

Now apply Theorem 4.1 to T. Note that to apply Theorem 4.1 to (5.1) on $(-\infty,0]$, the change of variable $t=f(x)=-x$ is made. This results in $m_-(\lambda)\to\infty$ as $\lambda\to-\infty$. Assume $v_-<v_+$ (the other cases are similar). From the equations (5.4) we can now draw the following conclusions. Note that a,b (c,d) have a continuous extension to the real axis except at poles of $m_+(m_-)$ (by the proof of Theorem (4.1)). Further, (4.9) implies $b(\lambda)>0$ on (v_+,∞) and $d(\lambda)<0$ on (v_-,∞). The argument below follows that of [10] in cases (1) and (2).

1. $v_+<\mu<\infty$. Then ρ is of class $C^{(1)}(v_+,\infty)$; further $\rho'_{11}(\mu)>0$, $\rho'_{22}(\mu)>0$ since $b(\mu)>0$, $d(\mu)<0$. Also rank $\rho'(\mu)=2$.

2. $v_-<\mu<v_+$. If μ is not a pole of m_+, then

$$\operatorname{Im}M(\mu)=\begin{pmatrix}-d/D & -ad/D\\ -ad/D & -a^2d/D\end{pmatrix}(\mu).$$

If μ is a pole of m_+, then as $\lambda\to\mu$

$$M(\lambda)\to\begin{pmatrix}0 & -1/2\\ -1/2 & m_-(\mu)\end{pmatrix},\qquad \operatorname{Im}M(\mu)\to\begin{pmatrix}0 & 0\\ 0 & -d(\mu)\end{pmatrix}.$$

Thus $\rho\in C^{(1)}(v_-,v_+)$, $\operatorname{rank}\rho'(\mu)=1$ on (v_-,v_+). Also $\rho'_{11}(\mu)>0$ except at a pole of m_+ in which case $\rho'_{11}(\mu)=0$, $\rho'_{22}(\mu)>0$; $\rho'_{22}(\mu)>0$ except at zeros of m_+.

3. $\mu<v_-$. m_+ and $m_-$are meromorphic on $\operatorname{Re}\lambda<v_-$. Thus $\operatorname{Im}M=0$ except at poles of M. These poles occur either where $a(\mu)=c(\mu)$ or where m_+ and m_- have a simultaneous pole. The poles of M, which are the eigenvalues of T, are bounded below since $m_+(\lambda)\to-\infty$, $m_-(\mu)\to\infty$ as $\lambda\to-\infty$. Thus v_- is the only possible accumulation point of these poles.

Finally we show at a pole μ_0 of M, $\mu_0<v_-$, the residue of M at μ_0 is of rank one. This means the eigenspace is of dimension one. If μ_0 is a pole of both m_+ and m_-, with residues σ_+ and σ_- respectively, then a calculation using (5.3) shows that the residue of M is

$$\lim_{\varepsilon\to0} i\varepsilon M(\mu_0+i\varepsilon)=\begin{pmatrix}0 & 0\\ 0 & \sigma_+\sigma_-/(\sigma_--\sigma_+)\end{pmatrix}.$$

If $m_-(\mu_0)=m_+(\mu_0)$ and $m_\pm$ are analytic at μ_0, then

$$m_\pm(\lambda)=a_0^\pm+a_1^\pm(\lambda-\mu_0)+\cdots$$

where all the $a_i^\pm$ are real. Further, $a_1^+>0$ since $\operatorname{Im}m_+(\lambda)>0$ on $\operatorname{Im}\lambda>$; similarly $a_1^-<0$. A calculation using (5.3) again gives the residue of M at μ_0 as

$$\begin{pmatrix}(a_1^+-a_1^-)^{-1} & (a_0^++a_0^-)/2(a_1^+-a_1^-)\\ (a_0^++a_0^-)/2(a_1^+-a_1^-) & a_0^+a_0^-/(a_1^+-a_1^-)\end{pmatrix}.$$

In both cases the residue has rank one.

The above calculations show a connection with recent work of R. Carmona [5]. Return to (5.3) and suppose on an interval I, one of m_+,m_-, say m_+, has $\operatorname{Im}m_+(\lambda)$ continuous on

$$I_\varepsilon=\{\lambda\,|\,\operatorname{Re}\lambda\in I,0<\operatorname{Im}\lambda\leqq\varepsilon\}$$

with a continuous extension to the closure of the above set. Further suppose on $\bar{I}_\varepsilon$, $\operatorname{Im} m_+(\lambda) > 0$. Hence on I_ε, $\operatorname{Im} m_+(\lambda)$ is bounded above and below by positive numbers. An examination of the formulas (5.4) shows that there is a constant k so that

$$|\operatorname{Im} m_{ij}(\lambda)| \leqq k$$

for $\lambda \in I_\varepsilon$, $i,j = 1,2$. This means that ρ_{ij} is absolutely continuous on I and has Lipschitz constant k. Thus the behavior of the potential at the other singular point cannot prevent the spectrum from being absolutely continuous on I. This phenomenon has been reported by R. Carmona [5] for a class of random polynomials which are of class $\mathscr{L}^p$, $p = 1$, at one singular endpoint.

6. The hydrogen atom. In this section we consider the energy operator of the hydrogen atom and show how the results of §§4 and 5 apply. The equation is [11, Chap. 10].

$$y'' + \left[\lambda + \frac{a}{x} - \frac{b}{x^2}\right] y = 0, \qquad 0 < x < \infty, \tag{6.1}$$

where $a > 0$ and $b = l(l+1) \geqq 0$. We split $0 < x < \infty$ at $x = 1$ and use any basis to define m_+ and m_-. First consider (6.1) on $1 \leqq x < \infty$. Then a/x is a v_1 type term of Theorem 4.1 and $-b/x^2$ is a v_3 type term. Thus m_+ behaves as follows. (i) On $0 < \lambda < \infty$, $\lim \operatorname{Im} m_+(\lambda + i\varepsilon)$ exists as $\varepsilon \downarrow 0$ and defines a continuous and positive function. (ii) On $\operatorname{Re}\lambda < 0$, m_+ is meromorphic and its poles are bounded below. If we appeal to the connection between oscillation theory and spectral theory [12, p. 163], we conclude that m_+ has an infinite number of poles on $(-\infty, 0)$ (thus they converge to 0) because (6.1) is oscillatory at infinity for $\lambda = 0$. The oscillation test $\int_1^x (as^{-1} - bs^{-2})\,dx \to \infty$ as $x \to \infty$ shows (6.1) is oscillatory at infinity for $\lambda = 0$ (cf. [12, p. 208]).

To consider m_-, we transform the singular point at $x = 0$. Let

$$y(x) = x^{1/2} z(t), \quad t = -\ln x, \quad 0 < x \leqq 1.$$

Then $x = 0$ is transformed to $t = \infty$, and a calculation shows that z satisfies the equation,

$$\ddot{z}(t) + \left[-\left(b + \frac{1}{4}\right) + ae^{-t} + \lambda e^{-2t}\right] z(t) = 0, \qquad 0 \leqq t < \infty. \tag{6.2}$$

By the Sturm comparison theorem (6.2) is nonoscillatory at infinity for every λ. Thus (6.1) is nonoscillatory at 0 for every λ. This means (cf. [12, p. 163]) any eigenvalue problem associated with (6.1) on $0 < x \leqq 1$ has as spectrum a sequence of eigenvalues increasing to infinity. Thus m_- is meromorphic on the complex plane with poles a sequence on the real axis increasing to infinity.

Thus by the argument of §5, we have for the spectral matrix of (6.1):

(i) On $(0, \infty)$: $\rho \in C^{(1)}(0, \infty)$; $\operatorname{rank} \rho'(\mu) = 1$; $\rho'_{11}(\mu) > 0$ except at a pole of m_- where $\rho'_{11} = 0$, $\rho'_{22} > 0$; $\rho'_{22}(\mu) > 0$ except at a zero of m_-.

(ii) On $(-\infty, 0)$: The sequence of (6.1) is purely discrete with spectrum bounded below. The eigenvalues cluster at 0. This latter point follows from considering a graph of $\operatorname{Re} m_+$, $\operatorname{Re} m_-$ on $(-\infty, 0)$. m_+ (which has negative residues) has a sequence of poles clustering at 0. m_- (which has positive residues) has a finite number of poles on $(-\infty, 0)$. The two graphs have a finitely number of intersections with cluster point zero. Alternatively, oscillation theory may be applied as in [12].

By the theory of §4 we could add perturbation terms to (6.1), such as $x \sin x^4$, without changing the basic conclusions above.

Acknowledgments. The authors gratefully acknowledge helpful discussions concerning this work with Professor S. G. Halvorsen of the University of Trondheim, Norway.

REFERENCES

[1] F. V. ATKINSON AND W. N. EVERITT, *Bounds for the point spectrum for a Sturm–Liouville equation*, Proc. Roy. Soc. Edinburgh, 80A (1978), pp. 57–66.

[2] M. BEN-ARTZI AND A. DEVINATZ, *Spectral and scattering theory for the adiabatic oscillator and related potentials*, J. Math. Phys. 20 (1979), pp. 594–607.

[3] M. BEN-ARTZI, *On the absolute continuity of Schrödinger operators with spherically symmetric, long range potentials*, I, J. Differential Equations 38 (1980), pp. 41–50.

[4] _____, *On the absolute continuity of Schrödinger operators with spherically symmetric, long range potentials*, II, J. Differential Equations 38 (1980), pp. 51–60.

[5] R. CARMONA, *One-dimensional Schrödinger operators with random or deterministic potentials: New spectral types*, J. Functional Anal., 51 (1983), pp. 229–258.

[6] J. CHANDHURI AND W. N. EVERITT, *On the spectrum of ordinary second order differential operators*, Proc. Roy. Soc. Edinburgh, 68A (1968), pp. 95–119.

[7] E. A. CODDINGTON AND N. LEVINSON, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.

[8] M. S. P. EASTHAM AND H. KALF, *Schrödinger-type operators with continuous spectra*, in Research Notes in Mathematics, vol. 65, Pitman, London, 1982.

[9] D. HINTON AND J. K. SHAW, *On the spectrum of a singular Hamiltonian system*, II, submitted.

[10] S. ITATSU AND H. KANETA, *Spectral matrices for first and second order self-adjoint ordinary differential operators with long range potentials*, Funkcialaj Ekvacioj, 24 (1981), pp. 23–45.

[11] E. MERZBACKER, *Quantum Mechanics*, John Wiley, New York, 1961.

[12] E. MÜLLER-PFEIFFER, *Spectral Theory of Ordinary Differential Operators*, Ellis Harwood, Chichester, 1981.

[13] P. A. REJTO, *On a theorem of Titchmarsh–Kodaira–Weidmann concerning absolutely continuous operators*, II, Indiana Univ. Math. J., 25 (1976), p. 629–658.

[14] P. A. REJTO AND K. SINHA, *Absolute continuity for a one-dimensional model of the Stark–Hamiltonian*, Helv. Phys. Acta, 49 (1976), pp. 389–413.

[15] P. A. REJTO, *An application of the third order* JWKB *approximation method to prove absolute continuity*, I, II, Helv. Phys. Acta, 50 91977), pp. 479–494; pp. 495–508.

[16] E. C. TITCHMARSH, *Eigenfunction Expansions Associated with Second-Order Differential Equations*, Part I, Oxford University Press, Oxford, 1982.

[17] J. WALTER, *Absolute continuity of the essential spectrum of* $-d^2/dt^2 + q(t)$ *without monotony of* q, Math. Z., 129 (1972), pp. 83–94.

[18] J. WEIDMANN, *Zur Spektraltheorie von Sturm–Liouville Operatoren*, Math. Z., 98 (1967), pp. 268–302.

SIAM J. MATH. ANAL.
Vol. 17, No. 1, January 1986

018

ASYMPTOTIC BEHAVIOR OF PERIODIC STRAIN STATES*

KENNETH B. HOWELL†

Abstract. The asymptotic behavior of the general periodic strain elastic state is discussed. The components of the elastic state are initially assumed to be bounded by an arbitrary polynomial. It is then shown that many of the components—for the case of plane strain, all of the components—can be approximated by second degree polynomials whose coefficients can be readily computed from data generally available in such problems. The error in using the approximation at different points in the elastic body is on the order of the reciprocal of any polynomial of the distance to the boundary of the body.

Consequences of these results are then discussed with regard to Saint-Venant's principle, the periodicity of solutions to periodic boundary value problems, the uniqueness of the solutions to periodic and "slightly periodic" boundary value problems, and various formulations of the theorem of work and energy.

1. Introduction. When dealing with problems involving infinite domains, one must have some concern about the asymptotic behavior of the functions "near infinity". Often, for example, it is desired—and assumed—that one or more of the functions and their derivatives rapidly and uniformly approach well defined (and computable) limits "at infinity". Muskhelishvili [6] and Gurtin and Sternberg [1] have shown the extent to which this type of asymptotic behavior can be expected in classical elastostatic problems on domains exterior to some compact set. (See also Knops and Payne [5, Chap. 6] for a discussion of similar problems on the whole- and half-space.) Unfortunately, it is not always clear that the asymptotic behavior one would desire or expect can be guaranteed on more complex domains. Indeed, in many cases just determining what behavior should be desired or expected is a significant problem in itself. The problem is often complicated by the fact that the asymptotic behavior may be strongly dependent on the direction along which "infinity" is approached. One class of problems in which this difficulty arises is the class of periodic and slightly periodic boundary value problems in elasticity. In this paper we shall examine the behavior of periodic strain states at great distances from the boundary of the elastic body. It shall be discovered—in Lemma 6.1 and Theorem 7.1—that certain (in some cases, all) components of the elastic state rapidly approach fixed values as the components are measured at increasing distances from the boundary of the domain. This "asymptotic state" is easily computed from the values of the displacement and traction on a portion of the boundary. In spite of the rather strong bounds which will be derived, the bounds initially assumed are quite weak—namely that the components of the elastic state are bounded by some arbitrary polynomial. After the derivation of these results, the implications of this asymptotic behavior will be discussed (in §§8 and 9) with regard to such issues as Saint-Venant's principle, the periodicity of solutions to periodic boundary value problems, and the uniqueness of solutions to slightly periodic boundary value problems. In addition, extensions of the theorem of work and energy will be discussed.

This is the second of two papers dealing with directionally dependent asymptotic behavior of biharmonic functions. In the first (Howell [4]) bounds were derived on the derivatives of the general biharmonic function based on bounds assumed for the original function. The major results of this first paper will be used extensively here, and are summarized here in §3.

*Received by the editors May 17, 1983, and in final revised form May 21, 1984.
† Department of Mathematics, University of Alabama in Huntsville, Huntsville, Alabama 35899.

2. Preliminaries: Elasticity. Whenever convenient, the points in k-dimensional Euclidean space will be identified with the vectors in $\mathbb{R}^k$ in the standard manner—i.e., through the agency of a suitably chosen Cartesian coordinate system. The induced orthonormal frame of vectors will be denoted by $\{\mathbf{e}^1, \mathbf{e}^2, \cdots, \mathbf{e}^k\}$. As usual, if $\mathbf{v}$ and $\mathbf{T}$ are, respectively, a vector and a (second rank) tensor, then v_i will denote $\mathbf{v}\cdot\mathbf{e}^i$ and T_{ij} will denote $\mathbf{e}^i\cdot\mathbf{T}\mathbf{e}^j$.

The (elastic) body will be denoted by $\mathscr{B}$ and its closure by $\mathrm{cl}\,\mathscr{B}$. It occupies an open connected subset of k-dimensional Euclidean space such that, if $\mathscr{D}$ is any ball of finite radius, then $\partial(\mathscr{B}\cap\mathscr{D})$ consists of a finite union of smooth $(k-1)$-dimensional manifolds with boundaries. In this paper, it will be assumed that the ambient space is either two- or three-dimensional and (within this range) k, the dimension of the space, will be arbitrary unless otherwise noted. The terms "volume" and "area" are to be interpreted accordingly. In particular, if $k=2$ then "volume" actually refers to area and "surface area" actually refers to arc length.

The outward pointing unit normal vector field on the boundary of $\mathscr{B}$ (or any subbody of $\mathscr{B}$ under discussion) will be denoted by $\mathbf{n}$.

The displacement, strain, and stress fields of an elastic state will be denoted by $\mathbf{u}$, $\mathbf{E}$, and $\mathbf{S}$, respectively, and are related throughout $\mathscr{B}$ by

$$\begin{aligned}\mathbf{E} &= \operatorname{sym}\nabla\mathbf{u},\\ \mathbf{S} &= \mathbf{C}[\mathbf{E}],\\ \operatorname{div}\mathbf{S}+\mathbf{b} &= \mathbf{0}\end{aligned}$$

where $\mathbf{C}$ is the elasticity tensor field on $\mathscr{B}$ and $\mathbf{b}$ denotes the body forces. If $\mathbf{n}$ is defined, $\mathbf{s}$ will denote the corresponding surface traction field, $\mathbf{Sn}$. Unless otherwise noted, the standard continuity and differentiability conditions assumed in classical treatments will be assumed here. The reader is reminded that $\mathbf{S}$ is a symmetric tensor field, that $\mathbf{C}[\mathbf{W}]$ vanishes whenever $\mathbf{W}$ is a skew tensor, and that if $\mathbf{E}$ vanishes in some region, then $\mathbf{u}$ is a rigid displacement in that region (i.e., $\nabla\mathbf{u}$ is a fixed skew tensor).

In this paper it shall always be assumed that the media composing $\mathscr{B}$ is homogeneous and isotropic. Thus, there will be two constants, μ and λ (the Lamé moduli) such that for any strain field, $\mathbf{E}$,

$$\mathbf{S}=\mathbf{C}[\mathbf{E}]=2\mu\mathbf{E}+\lambda(\operatorname{tr}\mathbf{E})\mathbf{Id}$$

where $\mathbf{Id}$ denotes the identity tensor. The corresponding displacement satisfies Navier's equation:

$$\mu\nabla\mathbf{u}+(\mu+\lambda)\nabla\operatorname{div}\mathbf{u}+\mathbf{b}=\mathbf{0}.$$

In addition, it is to be understood that $\mu(\lambda+2\mu)\neq 0$. Under these assumptions, it is well-known that, if $\mathbf{b}$ is both curl-free and divergence-free, then $\mathbf{u}$, $\mathbf{E}$, and $\mathbf{S}$ are biharmonic and infinitely differentiable on $\mathscr{B}$.

At times, additional assumptions will be made on the elasticity field, $\mathbf{C}$. One common assumption will be that $\mathbf{C}$ is positive definite. This means there exists a positive constant, c, such that given any symmetric tensor, $\mathbf{E}$, then

$$\mathbf{E}\cdot\mathbf{C}[\mathbf{E}]>c|\mathbf{E}|^2.$$

If $\mathscr{B}$ is two-dimensional, then $\mathbf{C}$ is positive definite if both $\mu>0$ and $\lambda+2\mu>0$. For three-dimensional bodies positive definiteness is equivalent to the Lamé moduli satisfying $\mu>0$ and $3\lambda+2\mu>0$. Occasionally, the slightly weaker assumption that $\mathbf{C}$ is strongly elliptic will be made. In this case the Lamé moduli satisfy $\mu(\lambda+2\mu)>0$, regardless of the dimension of the space.

For a given (mixed) boundary value problem, the Lamé moduli, μ and λ and the body force field, $\mathbf{b}$, are specified and the boundary of $\mathscr{B}$ is partitioned into two subsurfaces, $\mathscr{S}_1$ and $\mathscr{S}_2$, each of which is the domain of a predetermined vector field, $\hat{\mathbf{u}}$ and $\hat{\mathbf{s}}$, respectively. A solution to the boundary value problem consists of an elastic state, $(\mathbf{u}, \mathbf{E}, \mathbf{S})$ satisfying

$$\mathbf{u} = \hat{\mathbf{u}} \quad \text{on } \mathscr{S}_1,$$
$$\mathbf{s} = \hat{\mathbf{s}} \quad \text{on } \mathscr{S}_2.$$

If $\mathscr{S}_1$ consists of the entire boundary of $\mathscr{B}$, the problem is referred to as a (surface) displacement problem, while if $\mathscr{S}_2$ equals the boundary of $\mathscr{B}$ (up to a set of surface measure zero), then the problem is termed a (surface) traction problem. If the body is unbounded, one also usually desires that the elastic state at the point $\mathbf{x}$ behaves "reasonably" as $|\mathbf{x}|$ approaches infinity along one or more paths in $\mathscr{B}$. Precisely what is meant by "reasonable" and the extent to which "reasonable behavior" can be expected, depends on the particular problem at hand and is the main object of much of the study presented in this paper.

Given any mixed boundary value problem, there exists the corresponding null boundary value problem in which $\hat{\mathbf{u}}$, $\hat{\mathbf{s}}$, and $\mathbf{b}$ all vanish on their respective domains (μ, λ, $\mathscr{S}_1$, and $\mathscr{S}_2$ are as in the original problem). An obvious, but important, fact is that the difference between two solutions to the same mixed boundary value problem is a solution to the corresponding null boundary value problem.

Other types of boundary value problems can be described. In this paper the term "boundary value problem" implies that μ, λ, and $\mathbf{b}$ are prescribed and that if $(\mathbf{u}, \mathbf{E}, \mathbf{S})$ is the difference between any two solutions to the problem then $\mathbf{u} \cdot \mathbf{s}$ vanishes almost everywhere on the boundary of $\mathscr{B}$. It is trivial to verify that this includes the mixed boundary value problems described above. For this more general class of boundary value problems, the corresponding null boundary value problem is defined in the obvious manner.

The uniqueness of the solutions to various boundary value problems is discussed in the following two theorems. They will be used and refined later on in this paper.

THEOREM 2.1. *Let $\mathscr{B}$ be an unbounded body with positive definite elasticity field. Suppose that $(\mathbf{u}, \mathbf{E}, \mathbf{S})$ is the difference between two solutions to the same general boundary value problem on $\mathscr{B}$ and that*

$$\int_{\mathscr{B} \cap \partial \mathscr{D}_R} |\mathbf{u}|^2 da = O(R) \quad \text{as } R \to \infty$$

where $\mathscr{D}_R$ is the ball of radius R about the origin. Then $\mathbf{E}$ and $\mathbf{S}$ vanish on $\mathscr{B}$ and $\mathbf{u}$ is a rigid displacement.

THEOREM 2.2. *Let $\mathscr{B}$ be a two-dimensional unbounded body whose Lamé moduli satisfy*

$$\mu(2\mu + \lambda)(\mu + \lambda) \neq 0$$

and for which there exists a fixed vector, $\mathbf{v}^0$, such that

$$\mathbf{v}^0 \cdot \mathbf{n} > 0$$

almost everywhere (in the surface measure) on $\partial \mathscr{B}$. If $(\mathbf{u}, \mathbf{E}, \mathbf{S})$ is the difference between two solutions to the same traction problem on $\mathscr{B}$ such that

$$\int_{\mathscr{B} \cap \partial \mathscr{D}_R} |\nabla \mathbf{u}||\mathbf{S}| \, da = o(1) \quad \text{as } R \to \infty$$

where $\mathscr{D}_R$ is the ball of radius R about the origin, and

$$|\mathbf{S}(\mathbf{x})| = o(1) \quad \text{as } |\mathbf{x}| \to \infty,$$

then **E** *and* **S** *vanish on* $\mathscr{B}$ *and* **u** *is a rigid displacement.*

The above two theorems, in somewhat more general form, appear in Howell [2] (Theorems 2.1 and 5.2, respectively). In the same paper is a theorem very similar to Theorem 2.1 dealing with the displacement problem on bodies with a strongly elliptic elasticity (Theorem 4.1). It is not difficult, in fact, to show that all the results in the present paper involving a "general boundary value problem" on a body with "positive definite elasticity" will hold equally well for a "displacement problem" on a body with "strongly elliptic elasticity". This will be left to the interested reader.

3. Preliminaries: Asymptotic behavior of biharmonic functions. We are specifically concerned with the behavior of elastic states "near infinity". Since the components of the states which will be discussed are biharmonic, it is appropriate to quickly review the notation and major results reported in Howell [4] concerning the behavior of biharmonic functions "near infinity".

Let $\mathbf{x}$ be a fixed point in space, $\mathbf{v}$ a unit vector, and θ a scalar satisfying $0 < \theta < \pi/2$. The corresponding cone, $\mathscr{K} = \mathscr{K}(\mathbf{x}, \mathbf{v}, \theta)$, is the region in space given by

$$\mathscr{K} = \{\mathbf{y} : (\mathbf{y} - \mathbf{x}) \cdot \mathbf{v} < |\mathbf{y} - \mathbf{x}| \cos\theta\}.$$

This corresponds, of course, to the standard notion of a solid cone with vertex at $\mathbf{x}$, axis in the direction of $\mathbf{v}$, and whose sides make an angle of θ with the axis.

If, instead, one has a set of points in space, Σ, and a unit vector field on Σ, $\mathbf{v}(\mathbf{x})$, and a single fixed θ with $0 < \theta < \pi/2$, then the corresponding cone, $\mathscr{K} = \mathscr{K}(\Sigma, \mathbf{v}, \theta)$, is given by

$$\mathscr{K}(\Sigma, \mathbf{v}, \theta) = \bigcup_{\mathbf{x} \in \Sigma} \mathscr{K}(\mathbf{x}, \mathbf{v}(\mathbf{x}), \theta).$$

Finally, if n is a nonnegative integer, the nth subcone, $\mathscr{K}_n$, of $\mathscr{K}(\Sigma, \mathbf{v}, \theta)$ is defined by

$$\mathscr{K}_n = \mathscr{K}(\Sigma, \mathbf{v}, 2^{-n}\theta).$$

It may be noted that $\mathscr{K}_0 = \mathscr{K}$.

THEOREM 3.1. *Let Σ be a set of points in space, $\mathbf{v}$ a unit vector field on Σ, θ a constant with $0 < \theta \leqq \pi/2$, and $\mathscr{K} = \mathscr{K}(\Sigma, \mathbf{v}, \theta)$, and assume that*

$$\Sigma \subseteq \mathrm{cl}\mathscr{K}.$$

Let f be a positive-valued, locally integrable, nonincreasing function on $[0, \infty)$; let $\mathbf{m}$ be some fixed unit vector, and let ζ, ν, and c be real constants with c positive and $-1 < \nu$. Finally, let $\rho(\mathbf{x})$ denote the minimum distance from $\mathbf{x}$ to the closure of Σ.

i. *If $-1 < \nu \leqq 0$ and ϕ is a biharmonic function on $\mathscr{K}$ satisfying*

$$|\phi(\mathbf{x})| \leqq c(1 + |\mathbf{m} \cdot \mathbf{x}|)^{\nu} [\rho(\mathbf{x})]^{\zeta} f(\rho(\mathbf{x}))$$

for all $\mathbf{x}$ in $\mathscr{K}$, then

$$|\nabla\phi(\mathbf{x})| \leqq Bc(1 + |\mathbf{m} \cdot \mathbf{x}|)^{\nu} [\rho(\mathbf{x})]^{\zeta - 1} f\left(\frac{1}{2}\alpha\rho(\mathbf{x})\right)$$

for all $\mathbf{x}$ *in* $\mathscr{K}_1$, *where*

$$\alpha = \sin\left(\frac{1}{2}\theta\right), \qquad B = \frac{9}{2}\left[\frac{2}{\alpha}\right]^{|\zeta|+1} C_\nu$$

and C_ν *is a constant depending on* ν *only.*

ii. *If* $0 < \nu$ *and* ϕ *and* ψ *are biharmonic functions on* $\mathscr{K}$ *satisfying*

$$|\phi(\mathbf{x})| \leqq c(1+|\mathbf{m}\cdot\mathbf{x}|)^\nu [\rho(\mathbf{x})]^\zeta f(\rho(\mathbf{x})),$$
$$|\psi(\mathbf{x})| \leqq c\{[\rho(\mathbf{x})]^{\zeta+\nu} + (1+|\mathbf{m}\cdot\mathbf{x}|)^\nu [\rho(\mathbf{x})]^\zeta\} f(\rho(\mathbf{x}))$$

for all $\mathbf{x}$ *in* $\mathscr{K}$, *then*

$$|\nabla\phi(\mathbf{x})| \leqq Bc\{[\rho(\mathbf{x})]^{\zeta+\nu-1} + (1+|\mathbf{m}\cdot\mathbf{x}|)^\nu [\rho(\mathbf{x})]^{\zeta-1}\} f\left(\frac{1}{2}\alpha\rho(\mathbf{x})\right),$$
$$|\nabla\psi(\mathbf{x})| \leqq Mc\{[\rho(\mathbf{x})]^{\zeta+\nu-1} + (1+|\mathbf{m}\cdot\mathbf{x}|)^\nu [\rho(\mathbf{x})]^{\zeta-1}\} f\left(\frac{1}{2}\alpha\rho(\mathbf{x})\right)$$

for all $\mathbf{x}$ *in* $\mathscr{K}_1$, *where*

$$\alpha = \sin\left(\frac{1}{2}\theta\right),$$
$$B = \frac{9}{2^{\nu+1}}\left[\frac{2}{\alpha}\right]^{|\zeta|+\nu+1} C_\nu,$$
$$M = \frac{9[2^{|\zeta|+\nu} C_0 + 2^{|\zeta|} C_\nu]}{2^{|\zeta|+\nu+1}}$$

and C_0 *and* C_ν *are constants depending only on* ν.

4. Preliminaries: Periodicity. Let $\mathbf{p}$ be a fixed nonzero vector. A scalar-, vector-, or tensor-valued function, ϕ, with domain Ω is said to be periodic (with period $\mathbf{p}$) if both of the following hold:

1. $\mathbf{x}$ is in Ω if and only if $\mathbf{x}+\mathbf{p}$ is in Ω.
2. $\phi(\mathbf{x}+\mathbf{p}) = \phi(\mathbf{x})$ for every $\mathbf{x}$ in Ω.

For convenience, the following conventions will be implicit for the remainder of this paper:

1. The Cartesian coordinant system mentioned in the first paragraph of §2 is chosen so that

$$\mathbf{p} = p\mathbf{e}^1$$

where $p = |\mathbf{p}|$.

2. Unless otherwise stated, all periodic functions have the same period, $\mathbf{p}$.

A periodic boundary value problem is a boundary value problem in which the prescribed data ($\mathbf{b}$ and the boundary data) is given by periodic functions. An elastic state, $(\mathbf{u}, \mathbf{E}, \mathbf{S})$, on $\mathscr{B}$ may have periodic displacement, periodic strain, or periodic stress. Clearly, if the displacement is periodic so is the strain, and if the strain is periodic so is the stress. Likewise, if $\mathbf{C}$ (when restricted to the symmetric tensors) is invertible, then periodic stress implies periodic strain. Periodic strain, however, does not necessarily imply periodic displacement. A trivial example would be an elastic state in which $\mathbf{u}$ is a rigid displacement and $\mathbf{E}$ vanishes everywhere. Less trivial examples will be found in the next section. Perhaps even more disconcerting is an example of a solution to a periodic boundary value problem in which the strain is not periodic. Such an example may be found in Howell [3, §3].

A somewhat more general class of problems is the class of *slightly* periodic boundary value problems. A problem is said to be "slightly periodic" if its corresponding null boundary value problem—but not necessarily the original problem—is a periodic boundary value problem. An obvious example of a slightly periodic problem would be any traction problem on a half-space.

For many slightly periodic problems it will be convenient to define corresponding "period sections". Let $\mathbf{x}$ be any point in space and let $\mathscr{L}_{\mathbf{x}}^0$ be any plane through $\mathbf{x}$ which is not parallel to $\mathbf{p}$—that is, $(\mathbf{y}-\mathbf{x})\cdot\mathbf{p}$ is nonzero for every $\mathbf{y}\neq\mathbf{x}$ in $\mathscr{L}_{\mathbf{x}}^0$. $\mathscr{L}_{\mathbf{x}}$ will denote the intersection of $\mathscr{L}_{\mathbf{x}}^0$ with $\mathscr{B}$, while $\mathscr{P}_{\mathbf{x}}$ and $\mathscr{P}$ will both denote the set

$$\mathscr{P}=\{\mathbf{y}\in\mathscr{B}:\text{ for some } 0<\alpha<1,\ \mathbf{y}-\alpha\mathbf{p}\in\mathscr{L}_{\mathbf{x}}^0\}$$

$\mathscr{P}$ will also be referred to as a period section. It should be obvious that the boundary of $\mathscr{P}$ is the disjoint union of $\partial\mathscr{P}\cap\partial\mathscr{B}$ with the two parallel "faces" of $\mathscr{P}$, $\mathscr{L}_{\mathbf{x}}$ and $\mathscr{L}_{\mathbf{x}+\mathbf{p}}$. It should also be clear that if $\mathbf{n}$ is the outward normal vector field to $\mathscr{P}$ on $\partial\mathscr{P}$ and $\mathbf{y}$ is a point on $\mathscr{L}_{\mathbf{x}}$, then $\mathbf{n}(\mathbf{y})=-\mathbf{n}(\mathbf{y}+\mathbf{p})$.

Proofs of the following two theorems may be found in Howell [3, Lemma 4.1 and Thm. 3.1, respectively]. The first discusses the extent to which the displacement corresponding to a periodic strain state may, itself, fail to be periodic. The second theorem deals with the uniqueness of solutions to certain slightly periodic boundary value problems.

THEOREM 4.1. *Let* $(\mathbf{u},\mathbf{E},\mathbf{S})$ *be an elastic state on* $\mathscr{B}$ *with periodic strain,* $\mathbf{E}$. *Then, there is a fixed skew tensor,* $\overline{\mathbf{W}}$, *a constant,* κ, *and a fixed vector,* $\bar{\mathbf{u}}$, *such that*

$$\mathbf{p}\cdot\bar{\mathbf{u}}=0$$

and, for each $\mathbf{x}$ *in* $\mathscr{B}$,

$$\mathbf{u}(\mathbf{x}+\mathbf{p})-\mathbf{u}(\mathbf{x})=\overline{\mathbf{W}}\mathbf{x}+\kappa\mathbf{p}+\bar{\mathbf{u}}.$$

Furthermore, if $\mathbf{w}$ *is the rigid displacement given by*

$$\mathbf{w}(\mathbf{x})=p^{-2}[(\mathbf{p}\otimes\bar{\mathbf{u}})-(\bar{\mathbf{u}}\otimes\mathbf{p})]\mathbf{x}$$

and $\tilde{\mathbf{u}}$ *given by*

$$\tilde{\mathbf{u}}(\mathbf{x})=\mathbf{u}(\mathbf{x})+\mathbf{w}(\mathbf{x}),$$

then, for every $\mathbf{x}$ *in* $\mathscr{B}$,

$$\tilde{\mathbf{u}}(\mathbf{x}+\mathbf{p})-\tilde{\mathbf{u}}(\mathbf{x})=\overline{\mathbf{W}}\mathbf{x}+\kappa\mathbf{p}.$$

THEOREM 4.2. *Assume that the elasticity tensor is positive definite. Let* $(\mathbf{u},\mathbf{E},\mathbf{S})$ *be the difference between two solutions to the same general boundary value problem on* $\mathscr{B}$ *and assume that* $\mathbf{u}$ *is periodic. If, letting* $\mathscr{D}_R$ *be the circular cylinder of radius* R *about the* X_1*-axis,*

$$\int_{\mathscr{P}\cap\partial\mathscr{D}_R}|\mathbf{u}|^2\,da=O(R)\quad\text{as } R\to\infty, \tag{4.1}$$

then $\mathbf{E}$ *and* $\mathbf{S}$ *vanish on* $\mathscr{B}$ *and* $\mathbf{u}$ *is a rigid displacement.*

Henceforth, whenever $(\mathbf{u},\mathbf{E},\mathbf{S})$ is a periodic strain state, $\overline{\mathbf{W}}$, κ, and $\bar{\mathbf{u}}$ will denote, respectively, the skew tensor, constant, and vector whose existence is guaranteed by Theorem 4.1. In addition, if $\mathscr{B}$ is two-dimensional, then ω will denote the constant such that

$$\overline{\mathbf{W}}=\begin{pmatrix}0 & \omega\\ -\omega & 0\end{pmatrix}$$

while if $\mathscr{B}$ is three-dimensional, then ω_1, ω_2, and ω_3 will denote the constants such that

$$\overline{\mathbf{W}} = \begin{pmatrix} 0 & \omega_1 & \omega_2 \\ -\omega_1 & 0 & \omega_3 \\ -\omega_2 & -\omega_3 & 0 \end{pmatrix}.$$

It should be observed that the proof of Theorem 4.2, above, required a rather straightforward modification of the proof of Theorem 2.1. A similar modification can easily be made in the proof of Theorem 5.2 from Howell [2]. The resulting theorem (for two-dimensional problems) is given below. The details of the proof are left to the reader.

THEOREM 4.3. *Let $\mathscr{B}$ be a two-dimensional unbounded body whose Lamé moduli satisfy*

$$\mu(2\mu+\lambda)(\mu+\lambda) \neq 0$$

and for which there exists a fixed vector, $\mathbf{v}^0$, such that

$$\mathbf{v}^0 \cdot \mathbf{n} > 0$$

almost everywhere (in the surface measure) on $\partial\mathscr{B}$. Let $(\mathbf{u}, \mathbf{E}, \mathbf{S})$ be the difference between two solutions to the same slightly periodic traction problem on $\mathscr{B}$ such that

$$\int_{\mathscr{P}\cap\partial\mathscr{D}_R} |\nabla\mathbf{u}||\mathbf{S}|\, da = o(1) \quad \text{as } R \to \infty \tag{4.2}$$

where $\mathscr{D}_R$ is the cylinder of radius R about the X_1-axis, and

$$|\mathbf{S}(\mathbf{x})| = o(1) \quad \text{as } |\mathbf{x}| \to \infty \quad \text{on } \mathscr{P}. \tag{4.3}$$

If the stress field, $\mathbf{S}$, is periodic and the corresponding skew tensor, $\overline{\mathbf{W}}$, is the zero tensor, then $\mathbf{E}$ and $\mathbf{S}$ vanish on $\mathscr{B}$ and $\mathbf{u}$ is a rigid displacement.

For virtually every case of interest, Lemma 5.3 from Howell [3] can be employed to show that in Theorem 4.3, above, the assumption that $\overline{\mathbf{W}}$ vanish is unnecessary. In §6 of this paper, however, the asymptotic behavior of periodic strain states shall be studied in great detail. As a corollary of this study and Theorem 3.1, it shall be seen that, in many cases, uniqueness theorems comparable to the above two can be proven without assuming a periodic strain and with conditions (4.1), (4.2), and (4.3) replaced by much weaker assumptions.

5. Three important periodic strain states. Three special examples of periodic strain states will now be presented. To distinguish these states from the others, they shall be termed "base periodic strain states." While of some interest as counterexamples demonstrating the limitations of Theorems 4.2 and 4.3, the main reason for introducing these states—especially the first state below—will be their intimate relationship with the asymptotic behavior "near infinity" of other elastic states.

The base periodic strain state associated with two-dimensional periodic strain states. If $(\mathbf{u}, \mathbf{E}, \mathbf{S})$ is a two-dimensional periodic strain state with

$$\mathbf{u}(\mathbf{x}+\mathbf{p}) - \mathbf{u}(\mathbf{x}) = \overline{\mathbf{W}}\mathbf{x} + \kappa\mathbf{p} + \bar{\mathbf{u}},$$

then the associated base periodic strain state $(\mathbf{u}^*, \mathbf{E}^*, \mathbf{S}^*)$ is given by:

$$\mathbf{u}^*(x,y)=\frac{\omega}{2p}\begin{pmatrix}2xy\\-x^2\end{pmatrix}+\kappa\begin{pmatrix}x\\0\end{pmatrix}-\frac{\lambda}{2(2\mu+\lambda)p}\begin{pmatrix}0\\\omega y^2+2p\kappa y\end{pmatrix}+\left[\frac{\omega}{2}+\frac{\bar{u}_2}{p}\right]\begin{pmatrix}-y\\x\end{pmatrix},$$

$$\mathbf{E}^*=\operatorname{sym}\nabla\mathbf{u}^*,$$

$$\mathbf{S}^*=2\mu\mathbf{E}^*+\lambda(\operatorname{tr}\mathbf{E}^*)\,\mathbf{Id}=\frac{4\mu(\mu+\lambda)}{(2\mu+\lambda)p}\begin{pmatrix}\omega y+\kappa p & 0\\0 & 0\end{pmatrix}.$$

The following are easily verified:

1) $(\mathbf{u}^*,\mathbf{E}^*,\mathbf{S}^*)$ is a periodic strain state corresponding to zero body force (i.e. $\operatorname{div}\mathbf{S}^*=\mathbf{0}$).

2) $\mathbf{u}^*(\mathbf{x}+\mathbf{p})-\mathbf{u}^*(\mathbf{x})=\overline{\mathbf{W}}\mathbf{x}+\kappa\mathbf{p}+\bar{\mathbf{u}}=\mathbf{u}(\mathbf{x}+\mathbf{p})-\mathbf{u}(\mathbf{x})$.

3) $[\nabla\mathbf{u}^*(\mathbf{x})]\mathbf{p}=\overline{\mathbf{W}}\mathbf{x}+\kappa\mathbf{p}+\bar{\mathbf{u}}+\frac{1}{2}\omega\mathbf{p}$.

4) The surface traction, $\mathbf{s}^*=\mathbf{S}^*\mathbf{n}$, vanishes on any line parallel to the X_1-axis. Hence, $(\mathbf{u}^*,\mathbf{E}^*,\mathbf{S}^*)$ is a nontrivial solution to the null traction problem on any two-dimensional body whose boundary is parallel to the X_1-axis.

The base periodic strain state associated with three-dimensional periodic strain states. If $(\mathbf{u},\mathbf{E},\mathbf{S})$ is a three-dimensional periodic strain state with

$$\mathbf{u}(\mathbf{x}+\mathbf{p})-\mathbf{u}(\mathbf{x})=\overline{\mathbf{W}}\mathbf{x}+\kappa\mathbf{p}+\bar{\mathbf{u}},$$

then the associated base periodic strain state $(\mathbf{u}^*,\mathbf{E}^*,\mathbf{S}^*)$ is given by:

$$2p\mathbf{u}^*(x,y,z)=\begin{pmatrix}0\\-\omega_1\\-\omega_2\end{pmatrix}x^2+\begin{pmatrix}2\omega_1\\0\\-2\omega_3\end{pmatrix}xy+\begin{pmatrix}2\omega_2\\2\omega_3\\0\end{pmatrix}xz$$

$$+\frac{\lambda}{2(\mu+\lambda)}\left\{-2\kappa p\begin{pmatrix}0\\y\\z\end{pmatrix}+\begin{pmatrix}0\\-\omega_1\\\omega_2\end{pmatrix}y^2-\begin{pmatrix}0\\2\omega_2\\2\omega_1\end{pmatrix}yz+\begin{pmatrix}0\\\omega_1\\-\omega_2\end{pmatrix}z^2\right\}$$

$$+\begin{pmatrix}2\kappa p\\p\omega_1\\p\omega_2\end{pmatrix}x+\begin{pmatrix}-2\omega_3yz-p\omega_1y-p\omega_2z\\0\\0\end{pmatrix}-2p^{-1}[(\mathbf{p}\otimes\bar{\mathbf{u}})-(\bar{\mathbf{u}}\otimes\mathbf{p})]\mathbf{x},$$

$$\mathbf{E}^*=\operatorname{sym}\nabla\mathbf{u}^*,$$

$$\mathbf{S}^*(x,y,z)=2\mu\mathbf{E}^*+\lambda(\operatorname{tr}\mathbf{E}^*)\,\mathbf{Id}$$

$$=\begin{pmatrix}\frac{\mu(2\mu+3\lambda)}{(\mu+\lambda)p}[\omega_1y+\omega_2z+\kappa p] & 0 & \frac{-\omega_3y}{p}\\0 & 0 & 0\\\frac{-\omega_3y}{p} & 0 & 0\end{pmatrix}.$$

The following are easily verified:

1) $(\mathbf{u}^*,\mathbf{E}^*,\mathbf{S}^*)$ is a periodic strain state corresponding to zero body force (i.e., $\operatorname{div}\mathbf{S}^*=\mathbf{0}$).

2) $\mathbf{u}^*(\mathbf{x}+\mathbf{p})-\mathbf{u}^*(\mathbf{x})=\overline{\mathbf{W}}\mathbf{x}+\kappa\mathbf{p}+\bar{\mathbf{u}}=\mathbf{u}(\mathbf{x}+\mathbf{p})-\mathbf{u}(\mathbf{x})$.

3) $[\nabla\mathbf{u}^*(\mathbf{x})]\mathbf{p}=\overline{\mathbf{W}}\mathbf{x}+\kappa\mathbf{p}+\bar{\mathbf{u}}+\frac{1}{2}\omega_1p\mathbf{e}_2+\frac{1}{2}\omega_2p\mathbf{e}_3$.

4) $(\mathbf{u}^*,\mathbf{E}^*,\mathbf{S}^*)$ is a nontrivial solution to the null traction problem on any body whose boundary is parallel to the X_1-X_3 plane.

5) If $\omega_3=0$, $(\mathbf{u}^*,\mathbf{E}^*,\mathbf{S}^*)$ is a nontrivial solution to the null traction problem on any body whose boundary consists of straight lines parallel to the X_1-axis.

The alternate base periodic strain state associated with three-dimensional periodic strain states. If $(\mathbf{u}, \mathbf{E}, \mathbf{S})$ is a three-dimensional periodic strain state with

$$\mathbf{u}(\mathbf{x}+\mathbf{p})-\mathbf{u}(\mathbf{x})=\overline{\mathbf{W}}\mathbf{x}+\kappa\mathbf{p}+\bar{\mathbf{u}},$$

then the alternate associated base periodic strain state, $(\mathbf{u}^*, \mathbf{E}^*, \mathbf{S}^*)$ is given by

$$\begin{aligned}\mathbf{u}^*(x,y,z)=&\frac{\omega_1}{2p}\begin{pmatrix}2xy\\-x^2\\0\end{pmatrix}+\frac{\omega_2}{2p}\begin{pmatrix}2xz\\0\\-x^2\end{pmatrix}\\&+\frac{\omega_3}{p}\begin{pmatrix}0\\xz\\-xy\end{pmatrix}+\kappa\begin{pmatrix}x\\0\\0\end{pmatrix}-\frac{\lambda}{2p(2\mu+\lambda)}\begin{pmatrix}0\\\omega_1 y^2\\\omega_2 z^2\end{pmatrix}\\&-\frac{\lambda\kappa}{2(\mu+\lambda)}\begin{pmatrix}0\\y\\z\end{pmatrix}+\left[\frac{\omega_1}{2}+\frac{\bar{u}_2}{p}\right]\begin{pmatrix}-y\\x\\0\end{pmatrix}+\left[\frac{\omega_2}{2}+\frac{\bar{u}_3}{p}\right]\begin{pmatrix}-z\\0\\x\end{pmatrix},\end{aligned}$$

$$\mathbf{E}^*=\operatorname{sym}\nabla\mathbf{u}^*,$$

$$\mathbf{S}^*=2\mu\mathbf{E}^*+\lambda(\operatorname{tr}\mathbf{E}^*)\,\mathbf{Id}$$

$$=p^{-1}\begin{pmatrix}4\mu\dfrac{\mu+\lambda}{2\mu+\lambda}[\omega_1 y+\omega_2 z]+\mu\dfrac{2\mu+3\lambda}{\mu+\lambda}\kappa & \mu\omega_3 zp & -\mu\omega_3 y\\ \mu\omega_3 z & \dfrac{2\mu\lambda}{2\mu+\lambda}\omega_2 z & 0\\ -\mu\omega_3 y & 0 & \dfrac{2\mu\lambda}{2\mu+\lambda}\omega_1 y\end{pmatrix}.$$

The following are easily verified:

1) $(\mathbf{u}^*, \mathbf{E}^*, \mathbf{S}^*)$ is a periodic strain state corresponding to zero body force (i.e., $\operatorname{div}\mathbf{S}^*=\mathbf{0}$).
2) $\mathbf{u}^*(\mathbf{x}+\mathbf{p})-\mathbf{u}^*(\mathbf{x})=\overline{\mathbf{W}}\mathbf{x}+\kappa\mathbf{p}+\bar{\mathbf{u}}=\mathbf{u}(\mathbf{x}+\mathbf{p})-\mathbf{u}(\mathbf{x})$.
3) $[\nabla\mathbf{u}^*(\mathbf{x})]\mathbf{p}=\overline{\mathbf{W}}\mathbf{x}+\kappa\mathbf{p}+\bar{\mathbf{u}}+\frac{1}{2}\omega_1 p\mathbf{e}_2+\frac{1}{2}\omega_2 p\mathbf{e}_3$.
4) If $\omega_2=\omega_3=0$, $(\mathbf{u}^*, \mathbf{E}^*, \mathbf{S}^*)$ is a nontrivial solution to the null traction problem on any body whose boundary is parallel to the X_1-X_3 plane.
5) If $\omega_1=\omega_2=0$, $(\mathbf{u}^*, \mathbf{E}^*, \mathbf{S}^*)$ is a nontrivial solution to the null traction problem on any body whose boundary is a right circular cylinder centered on the X_1-axis.
6) If $\omega_1=\omega_2=\omega_3=0$, $(\mathbf{u}^*, \mathbf{E}^*, \mathbf{S}^*)$ is a nontrivial solution to the null traction problem on any body whose boundary consists of straight lines parallel to the X_1-axis.

6. The major lemma.

LEMMA 6.1. *Suppose* $\mathscr{K}=\mathscr{K}(\Sigma, \mathbf{v}, \theta)$ *is a nontrivial cone* $(\theta\neq 0)$ *satisfying the following conditions*:

i. Σ *is a nontrivial closed subset of* $\partial\mathscr{K}$;

ii. *for each real* η *and each nonnegative integer,* m, $\mathbf{x}+\eta\mathbf{e}^1$ *is contained in* $\mathscr{K}_m=\mathscr{K}(\Sigma, \mathbf{v}, 2^{-m}\theta)$ *whenever* $\mathbf{x}$ *is a point in* $\mathscr{K}_1$.

Let $(\mathbf{u}, \mathbf{E}, \mathbf{S})$ *be a periodic strain state on* $\mathscr{K}$ *(with period* $\mathbf{p}$*) corresponding to a body force with vanishing curl and divergence, and suppose that for some fixed* $\beta \geqq 2$*, either*

$$\mathbf{u}(\mathbf{x}) = O\left(|\mathbf{x}|^{\beta}\right) \quad \text{as } |\mathbf{x}| \to \infty \quad \text{on } \mathscr{K} \tag{6.1}$$

or

$$\mathbf{E}(\mathbf{x}) = O\left(|\mathbf{x}|^{\beta-1}\right) \quad \text{as } |\mathbf{x}| \to \infty \quad \text{on } \mathscr{K}_1. \tag{6.2}$$

Then, given any integer, $n \geqq 2$,

$$[\nabla \mathbf{u}(\mathbf{x})]\mathbf{p} - [\overline{\mathbf{W}}\mathbf{x} + \kappa \mathbf{p} + \bar{\mathbf{u}} + \mathbf{h}] = O(\rho^{\beta-n}) \quad \text{as } \rho \to \infty \quad \text{on } \mathscr{K}_n$$

where $\rho = \rho(\mathbf{x})$ *denotes the distance from* $\mathbf{x}$ *to* Σ *and*

$$\mathbf{h} = \begin{cases} \frac{1}{2} p \omega \mathbf{e}^2 & \text{if } \mathbf{E} \text{ is a plane strain,} \\ \frac{1}{2} p \omega_1 \mathbf{e}^2 + \frac{1}{2} p \omega_2 \mathbf{e}^3 & \text{otherwise.} \end{cases}$$

Proof. It should be noted that (6.1) and (6.2) both imply

$$\mathbf{E}(\mathbf{x}) = O(\rho^{\beta-1}) \quad \text{as } \rho \to \infty \quad \text{on } \mathscr{K}_1. \tag{6.3}$$

To see this, first observe that if (6.1) holds, then there must exist some nonincreasing, locally integrable function, $f(\rho)$, such that

$$|\mathbf{u}(\mathbf{x})| \leqq \left[\rho^{\beta} + \left(1 + |x_1|\right)^{\beta}\right] f(\rho)$$

for each $\mathbf{x}$ in $\mathscr{K}$. But, since $\mathbf{E}$ is the symmetric gradient of $\mathbf{u}$, Theorem 3.1 (with $\nu = 0$ and $\zeta = \beta$) can be employed to show that

$$|\mathbf{E}(\mathbf{x})| \leqq M\left[\rho^{\beta-1} + \left(1 + |x_1|\right)^{\beta} \rho^{-1}\right] f\left(\frac{1}{2}\alpha\rho\right)$$

for each $\mathbf{x}$ in $\mathscr{K}_1$ (M and α are constants from Theorem 3.1). Thus, by the geometry of the period sections,

$$\mathbf{E}(\mathbf{x}) = O(\rho^{\beta-1}) \quad \text{as } \rho \to \infty \quad \text{on } \mathscr{K}_1 \cap \mathscr{P} \tag{6.4}$$

where $\mathscr{P}$ is any period section. In a much more obvious manner, (6.4) also follows from (6.2). (6.3) now follows immediately since $\mathbf{E}$ is periodic on $\mathscr{K}_1$.

Now, let $(\tilde{\mathbf{u}}, \tilde{\mathbf{E}}, \tilde{\mathbf{S}})$ be the elastic state defined by

$$\tilde{\mathbf{u}} = \mathbf{u} - \mathbf{u}^*$$

where $(\mathbf{u}^*, \mathbf{E}^*, \mathbf{S}^*)$ is any of the corresponding base periodic strain states associated with $(\mathbf{u}, \mathbf{E}, \mathbf{S})$. Clearly, $(\tilde{\mathbf{u}}, \tilde{\mathbf{E}}, \tilde{\mathbf{S}})$ is a periodic displacement state on $\mathscr{K}_1$ which also satisfies (6.3), that is

$$\tilde{\mathbf{E}}(\mathbf{x}) = O(\rho^{\beta-1}) \quad \text{as } \rho \to \infty \quad \text{on } \mathscr{K}_1. \tag{6.5}$$

Employing Theorem 3.1 again (this time with $\nu = 0$ and $\zeta = \beta - 1$), it can be seen that (6.5) implies

$$\nabla \tilde{\mathbf{E}}(\mathbf{x}) = O(\rho^{\beta-2}) \quad \text{as } \rho \to \infty \quad \text{on } \mathscr{K}_2.$$

But it is easily checked that

$$\tilde{u}_{i,jk} = \tilde{E}_{ij,k} - \tilde{E}_{ik,j} + \tilde{E}_{jk,i}$$

and, thus,

$$\nabla\nabla\tilde{\mathbf{u}}(\mathbf{x}) = O(\rho^{\beta-2}) \quad \text{as } \rho\to\infty \quad \text{on } \mathscr{K}_2.$$

By this, the periodicity of $\tilde{\mathbf{u}}$, condition ii., and basic calculus:

$$\begin{aligned}[\nabla\tilde{\mathbf{u}}(\mathbf{x})]\mathbf{p} &= [\nabla\tilde{\mathbf{u}}(\mathbf{x})]\mathbf{p} - [\tilde{\mathbf{u}}(\mathbf{x}+\mathbf{p}) - \tilde{\mathbf{u}}(\mathbf{x})] \\ &= \int_{\sigma=0}^{p} [\nabla\tilde{\mathbf{u}}(\mathbf{x}) - \nabla\tilde{\mathbf{u}}(\mathbf{x}+\sigma\mathbf{e}^1)]\mathbf{e}^1\, d\sigma \\ &= -\int_{\sigma=0}^{p}\int_{\tau=0}^{\sigma} [\nabla\tilde{\mathbf{u}}_{,1}(\mathbf{x}+\tau\mathbf{e}^1)]\mathbf{e}^1\, d\tau\, d\sigma \\ &= O(\rho^{\beta-2}) \quad \text{as } \rho\to\infty \quad \text{on } \mathscr{K}_2,\end{aligned}$$

which, of course, is equivalent to

$$\tilde{\mathbf{u}}_{,1}(\mathbf{x}) = O(\rho^{\beta-2}) \quad \text{as } \rho\to\infty \quad \text{on } \mathscr{K}_2 \tag{6.6}$$

since $\mathbf{p} = p\mathbf{e}^1$.

Theorem 3.1 can now be applied to $\mathbf{u}_{,1}$ using the bounds indicated by (6.6). The result is that

$$\nabla\tilde{\mathbf{u}}_{,1}(\mathbf{x}) = O(\rho^{\beta-3}) \quad \text{as } \rho\to\infty \quad \text{on } \mathscr{K}_3.$$

The above integration is still valid but now yields

$$\begin{aligned}[\nabla\tilde{\mathbf{u}}(\mathbf{x})]\mathbf{p} &= -\int_{\sigma=0}^{p}\int_{\tau=0}^{\sigma} [\nabla\tilde{\mathbf{u}}_{,1}(\mathbf{x}+\tau\mathbf{e}^1)]\mathbf{e}^1\, d\tau\, d\sigma \\ &= O(\rho^{\beta-3}) \quad \text{as } \rho\to\infty \quad \text{on } \mathscr{K}_3.\end{aligned}$$

Thus,

$$[\nabla\tilde{\mathbf{u}}(\mathbf{x})]\mathbf{p} = O(\rho^{\beta-2}) \quad \text{as } \rho\to\infty \quad \text{on } \mathscr{K}_2 \tag{6.7}$$

implies

$$[\nabla\tilde{\mathbf{u}}(\mathbf{x})]\mathbf{p} = O(\rho^{\beta-3}) \quad \text{as } \rho\to\infty \quad \text{on } \mathscr{K}_3. \tag{6.8}$$

Repeating the arguments which led from (6.7) to (6.8), one quickly discovers that (6.8) implies

$$[\nabla\tilde{\mathbf{u}}(\mathbf{x})]\mathbf{p} = O(\rho^{\beta-4}) \quad \text{as } \rho\to\infty \quad \text{on } \mathscr{K}_4$$

which, in turn, leads to

$$[\nabla\tilde{\mathbf{u}}(\mathbf{x})]\mathbf{p} = O(\rho^{\beta-5}) \quad \text{as } \rho\to\infty \quad \text{on } \mathscr{K}_5$$

and so forth. Taking the inductive leap, one obtains

$$[\nabla\tilde{\mathbf{u}}(\mathbf{x})]\mathbf{p} = O(\rho^{\beta-n}) \quad \text{as } \rho\to\infty \quad \text{on } \mathscr{K}_n$$

for every integer, $n \geqq 2$. This essentially completes the proof since by the definition of $\tilde{\mathbf{u}}$

$$[\nabla\tilde{\mathbf{u}}(\mathbf{x})]\mathbf{p} = [\nabla\mathbf{u}(\mathbf{x})]\mathbf{p} - [\nabla\mathbf{u}^*(\mathbf{x})]\mathbf{p}$$

and, as was observed in §5,

$$[\nabla\mathbf{u}^*(\mathbf{x})]\mathbf{p} = \overline{\mathbf{W}}\mathbf{x} + \kappa\mathbf{p} + \bar{\mathbf{u}} + \mathbf{h}$$

for any of the base periodic strain states associated with $(\mathbf{u}, \mathbf{E}, \mathbf{S})$. □

It should be observed that, often, $\mathscr{K}=\mathscr{K}_1=\mathscr{K}_2$, etc. In these cases, the above results are quite strong. Examples of such situations would be problems involving half-spaces and bodies exterior to cylinders (more generally, periodic fractional spaces and periodic exterior bodies, see Howell [4, §7]).

In the next section Lemma 6.1 will prove extremely useful in establishing completely the asymptotic behavior of periodic plane strain states. A somewhat simpler application of Lemma 6.1 can be found in the next corollary.

COROLLARY 6.2. *Let $\mathscr{K}$ be as in Lemma* 6.1 *and let f denote a locally integrable, nonincreasing function on $[0,\infty)$ such that*

$$f(\rho)=o(1) \quad \text{as } \rho\to\infty.$$

Suppose that $(\mathbf{u},\mathbf{E},\mathbf{S})$ *is a periodic strain state on a body containing $\mathscr{K}$ and suppose that either*

$$|\mathbf{u}(\mathbf{x})|\leqq\left[\rho+(1+|x_1|)\right]f(\rho) \tag{6.9}$$

for every $\mathbf{x}$ *in $\mathscr{K}$ or that*

$$|\mathbf{S}(\mathbf{x})|\leqq f(\rho) \tag{6.10}$$

for every $\mathbf{x}$ *in $\mathscr{K}_1$. Then* $|\overline{\mathbf{W}}|=\kappa=0$.

Proof. Using the invertibility of the relationship between the strain and the stress, and arguments similar to those used in the proof of Lemma 6.1, it can be shown that either (6.9) or (6.10) implies that

$$\nabla\nabla\mathbf{u}(\mathbf{x})=O(\rho^{-1}) \quad \text{as } \rho\to\infty \quad \text{on } \mathscr{K}_3$$

and

$$E_{11}(\mathbf{x})=o(1) \quad \text{as } \rho\to\infty \quad \text{on } \mathscr{K}_2.$$

It then follows that

$$\begin{aligned}\overline{\mathbf{W}}&=\nabla\mathbf{u}(\mathbf{x}+\mathbf{p})-\nabla\mathbf{u}(\mathbf{x})\\&=\int_{\sigma=0}^{p}\nabla\mathbf{u}_{,1}(\mathbf{x}+\sigma\mathbf{e}^1)\,d\sigma=O(\rho^{-1}) \quad \text{as } \rho\to\infty \quad \text{on } \mathscr{K}_3\end{aligned}$$

which is possible only if $\overline{\mathbf{W}}$ is the zero tensor.

Now, since $\overline{\mathbf{W}}$ is the zero tensor, Lemma 6.1 implies that

$$[\nabla\mathbf{u}(\mathbf{x})]\mathbf{p}-[\kappa\mathbf{p}+\mathbf{u}]=O(\rho^{-2}) \quad \text{as } \rho\to\infty \quad \text{on } \mathscr{K}_4.$$

Here, we have also used the fact that $\mathbf{h}$ vanishes if $\overline{\mathbf{W}}$ vanishes and the fact that by either (6.9) or (6.10), $\beta=2$. Taking the first component, rearranging slightly, and recalling that $E_{11}=u_{1,1}$ gives

$$\kappa p=pE_{11}(\mathbf{x})+O(\rho^{-2}) \quad \text{as } \rho\to\infty \quad \text{on } \mathscr{K}_4$$

which, since $E_{11}(\mathbf{x})$ also vanishes as $\rho\to\infty$, forces κ to be zero. □

7. Asymptotic behavior of periodic plane strain states. In this section and the next, attention will be restricted to plane strain states on bodies containing the upper half-plane, $\{\mathbf{x}=(x,y):y>0\}$. For convenience, the notation (x,y) rather than (x_1,x_2) will be adopted for the components of $\mathbf{x}$. It should be clear that the upper half-plane can be viewed as the cone $\mathscr{K}=\mathscr{K}(\Sigma_0,\mathbf{e}^2,\theta)$, where Σ_0 is the X_1-axis and θ is arbitrary. Thus, the results from the previous section can be applied. Let us also, at this time, take

note of the following two observations:

1. $\rho(\mathbf{x})$, the distance between Σ_0 and $\mathbf{x}=(x,y)$, is given by $|y|$.
2. For any nonnegative integer, n, $\mathscr{K}_n=\mathscr{K}$.

It will be shown that at great distances from the boundary, any periodic plane strain state, $(\mathbf{u},\mathbf{E},\mathbf{S})$, can be closely approximated once four constants corresponding to the state have been determined. Not surprisingly, two of these constants are ω and κ. The other two constants are the components of the "negative mean surface traction," $\bar{\mathbf{s}}$, defined by

$$\bar{\mathbf{s}}=\lim_{y\to-\infty}\frac{-1}{p}\int_{\mathscr{S}_y}\mathbf{S}\mathbf{n}\,da \tag{7.1}$$

where, for any period section, $\mathscr{P}$,

$$\mathscr{S}_y=\partial\big(\mathscr{B}\cap\{(x,\hat{y}):\hat{y}\geqq y\}\big)\cap\mathrm{cl}\,\mathscr{P}.$$

It will be seen that the negative mean surface traction does always exist.

Let $(\mathbf{u},\mathbf{E},\mathbf{S})$ be a periodic strain state with corresponding negative mean surface traction $\bar{\mathbf{s}}=(\bar{s}_1,\bar{s}_2)$ and corresponding base periodic strain state $(\mathbf{u}^*,\mathbf{E}^*,\mathbf{S}^*)$ (as defined in §5 for plane strains). The "asymptotic state", $(\mathbf{u}^0,\mathbf{E}^0,\mathbf{S}^0)$, corresponding to $(\mathbf{u},\mathbf{E},\mathbf{S})$ is given by the following formulas:

$$\begin{aligned}\mathbf{u}^0(x,y)&=\mathbf{u}^*(x,y)+\frac{\lambda}{\mu(2\mu+\lambda)}\begin{pmatrix}[2\mu+\lambda]\bar{s}_1y\\ \mu\bar{s}_2y\end{pmatrix}\\&=\frac{\omega}{2p}\begin{pmatrix}2xy\\ -x^2\end{pmatrix}+\frac{\lambda}{2p(2\mu+\lambda)}\begin{pmatrix}0\\ \omega y^2+2p\kappa y\end{pmatrix}\\&\quad+\kappa\begin{pmatrix}x\\0\end{pmatrix}+\left[\frac{\omega}{2}+\frac{\bar{u}_2}{p}\right]\begin{pmatrix}-y\\ x\end{pmatrix}+\frac{1}{\mu(2\mu+\lambda)}\begin{pmatrix}[2\mu+\lambda]\bar{s}_1y\\ \mu\bar{s}_2y\end{pmatrix},\end{aligned} \tag{7.2}$$

$$\mathbf{E}^0=\operatorname{sym}\nabla\mathbf{u}^0=\mathbf{E}^*+\frac{1}{2\mu(2\mu+\lambda)}\begin{pmatrix}0&(2\mu+\lambda)\bar{s}_1\\ (2\mu+\lambda)\bar{s}_1&2\mu\bar{s}_2\end{pmatrix}, \tag{7.3}$$

$$\mathbf{S}^0(x,y)=2\mu\mathbf{E}^0+\lambda(\operatorname{tr}\mathbf{E}^0)\,\mathbf{Id}=\begin{pmatrix}A[\omega y+\kappa p]+B\bar{s}_2&\bar{s}_1\\ \bar{s}_1&\bar{s}_2\end{pmatrix}, \tag{7.4}$$

where

$$A=\frac{4\mu(\mu+\lambda)}{p(2\mu+\lambda)},\qquad B=\frac{\lambda}{2\mu+\lambda}.$$

It is easily verified that $(\mathbf{u}^0,\mathbf{E}^0,\mathbf{S}^0)$ is a periodic strain state corresponding to zero body force (i.e. $\operatorname{div}\mathbf{S}^0=\mathbf{0}$) and that $\mathbf{u}-\mathbf{u}^0$ is a periodic displacement. The next theorem establishes much more, namely that $(\mathbf{u}^0,\mathbf{E}^0,\mathbf{S}^0)$ closely approximates $(\mathbf{u},\mathbf{E},\mathbf{S})$ at great distances from the boundary of the elastic body.

THEOREM 7.1. *Suppose that $\mathscr{B}$ is a two-dimensional elastic body containing the upper half plane, $\{(x,y):y>0\}$, and that $(\mathbf{u},\mathbf{E},\mathbf{S})$ is a periodic strain state on $\mathscr{B}$ corresponding to zero body force. Assume, also, that for some arbitrary real β and some period section, $\mathscr{P}$, either*

$$\mathbf{u}(\mathbf{x})=O\big(|\mathbf{x}|^{\beta}\big)\quad \text{as } |\mathbf{x}|\to\infty\quad \text{on } \mathscr{P}$$

or

$$\mathbf{E}(\mathbf{x}) = O\left(|\mathbf{x}|^{\beta}\right) \quad \text{as } |\mathbf{x}| \to \infty \quad \text{on } \mathscr{P}.$$

Then the negative mean surface traction, $\bar{\mathbf{s}}$, *exists. Moreover, there is a fixed vector,* $\mathbf{v}^0$, *such that, for any positive* γ,

$$\mathbf{u}(x,y) - \mathbf{u}^0(x,y) - \mathbf{v}^0 = 0(y^{-\gamma}) \quad \text{as } y \to \infty, \tag{7.5}$$

$$\mathbf{E}(x,y) - \mathbf{E}^0(x,y) = O(y^{-\gamma}) \quad \text{as } y \to \infty, \tag{7.6}$$

$$\mathbf{S}(x,y) - \mathbf{S}^0(x,y) = O(y^{-\gamma}) \quad \text{as } y \to \infty, \tag{7.7}$$

where $(\mathbf{u}^0, \mathbf{E}^0, \mathbf{S}^0)$ *is the asymptotic state corresponding to* $(\mathbf{u}, \mathbf{E}, \mathbf{S})$ *given by formulas* (7.2), (7.3), *and* (7.4).

Proof. Let $(\mathbf{u}^*, \mathbf{E}^*, \mathbf{S}^*)$ be the base periodic strain state corresponding to $(\mathbf{u}, \mathbf{E}, \mathbf{S})$ and let $(\tilde{\mathbf{u}}, \tilde{\mathbf{E}}, \tilde{\mathbf{S}})$ be the elastic state on defined by

$$\tilde{\mathbf{u}}(\mathbf{x}) = \mathbf{u}(\mathbf{x}) - \mathbf{u}^*(\mathbf{x}).$$

It is easily checked that $(\tilde{\mathbf{u}}, \tilde{\mathbf{E}}, \tilde{\mathbf{S}})$ is a periodic displacement state on $\mathscr{B}$ corresponding to zero body force and, by Lemma 6.1,

$$[\nabla \tilde{\mathbf{u}}(x,y)]\mathbf{p} = O(y^{-\gamma-1}) \quad \text{as } y \to \infty$$

for all positive γ. By Theorem 3.1, then,

$$\tilde{u}_{i,jk}(x,y) = O(y^{-\gamma-2}) \quad \text{as } y \to \infty \tag{7.8}$$

for all possible i, j, and k except $i=j=2$. However, since the body force is zero, the two-dimensional Navier equation can be written

$$\tilde{u}_{1,11} + \tilde{u}_{1,22} + \frac{\mu+\lambda}{\mu}(\tilde{u}_{1,11} + \tilde{u}_{2,12}) = 0,$$

$$\tilde{u}_{2,11} + \tilde{u}_{2,22} + \frac{\mu+\lambda}{\mu}(\tilde{u}_{1,12} + \tilde{u}_{2,22}) = 0$$

which can be solved for $\tilde{u}_{1,22}$ and $\tilde{u}_{2,22}$ in terms of the other second derivatives of $\tilde{\mathbf{u}}$. But, it is for these other second derivatives that (7.8) holds. Thus,

$$\nabla\nabla\tilde{\mathbf{u}}(x,y) = O(y^{-\gamma-2}) \quad \text{as } y \to \infty \tag{7.9}$$

for any positive γ.

Now, choose any pair of points (x,y) and $(\bar{x},\bar{y})$ with $|x-\bar{x}| \leqq p$ and $0 < y \leqq \bar{y}$. Integrating $\nabla\nabla\tilde{\mathbf{u}}$ along a straight line path between (x,y) and $(\bar{x},\bar{y})$ leads to

$$\nabla\tilde{\mathbf{u}}(x,y) - \nabla\tilde{\mathbf{u}}(\bar{x},\bar{y}) = O(y^{-\gamma-1}) \quad \text{as } y \to \infty$$

for any positive γ. Straightforward "Cauchy sequence" arguments shows that this implies the existence of a fixed tensor, $\hat{\mathbf{T}}$, such that

$$\nabla\tilde{\mathbf{u}}(x,y) - \hat{\mathbf{T}} = O(y^{-\gamma-1}) \quad \text{as } y \to \infty.$$

Using either the periodicity of $\tilde{\mathbf{u}}$ or the asymptotic behavior of $\nabla\nabla\mathbf{u}(x,y)$ as $y \to \infty$, it is easily seen that $\hat{\mathbf{T}}$ is independent of the choice of x.

Using $\nabla \mathbf{u}(x,y)-\hat{\mathbf{T}}$ instead of $\nabla \nabla \mathbf{u}(x,y)$, the above succession of integration and "Cauchy sequence" arguments proves that there is a fixed vector, $\mathbf{v}^0$, such that

$$\tilde{\mathbf{u}}(\mathbf{x})-\hat{\mathbf{T}}\mathbf{x}-\mathbf{v}^0=O(y^{-\gamma}) \quad \text{as } y\to\infty \tag{7.10}$$

where $\mathbf{x}=(x,y)$ and γ is any positive real constant.

Letting $(\hat{\mathbf{u}},\hat{\mathbf{E}},\hat{\mathbf{S}})$ be the elastic state given by

$$\begin{aligned}\hat{\mathbf{u}}(\mathbf{x})&=\hat{\mathbf{T}}\mathbf{x},\\ \hat{\mathbf{E}}&=\operatorname{sym}\nabla\hat{\mathbf{u}},\\ \hat{\mathbf{S}}&=2\mu\hat{\mathbf{E}}+\lambda(\operatorname{tr}\hat{\mathbf{E}})\,\mathbf{Id}\end{aligned}$$

and recalling the definition of $\tilde{\mathbf{u}}$, it becomes clear that (7.10) is equivalent to

$$\mathbf{u}(\mathbf{x})-\mathbf{u}^*(\mathbf{x})-\hat{\mathbf{u}}(\mathbf{x})-\mathbf{v}^0=O(y^{-\gamma}) \quad \text{as } y\to\infty \tag{7.11}$$

which by Theorem 3.1, implies

$$\mathbf{E}(\mathbf{x})-\mathbf{E}^*(\mathbf{x})-\hat{\mathbf{E}}=O(y^{-\gamma}) \quad \text{as } y\to\infty \tag{7.12}$$

and

$$\mathbf{S}(\mathbf{x})-\mathbf{S}^*(\mathbf{x})-\hat{\mathbf{S}}=O(y^{-\gamma}) \quad \text{as } y\to\infty \tag{7.13}$$

where $\mathbf{x}=(x,y)$ and γ is any positive real constant. Comparing the last three expressions with (7.5) through (7.7), along with (7.2) through (7.4), leads to the realization that the theorem will be proven once $\bar{\mathbf{s}}$ has been shown to exist and once it has been shown that

$$\hat{\mathbf{T}}\mathbf{x}=\frac{1}{\mu(2\mu+\lambda)}\begin{pmatrix}[2\mu+\lambda]\bar{s}_1 y\\ \mu\bar{s}_2 y\end{pmatrix} \tag{7.14}$$

for each $\mathbf{x}=(x,y)$. It should be observed that (7.14) certainly holds if both $\hat{T}_{i1}=0$ and $\hat{S}_{i2}=\bar{s}_i$ for $i=1$ and 2.

First, consider the fixed tensor $\hat{\mathbf{T}}$. By the periodicity of $\tilde{\mathbf{u}}$ and (7.10)

$$\hat{\mathbf{T}}\mathbf{p}=[\tilde{\mathbf{u}}(\mathbf{x})-\hat{\mathbf{T}}\mathbf{x}]-[\tilde{\mathbf{u}}(\mathbf{x}+\mathbf{p})-\hat{\mathbf{T}}\mathbf{x}-\hat{\mathbf{T}}\mathbf{p}]=O(y^{-\gamma}) \quad \text{as } y\to\infty$$

for each $\mathbf{x}=(x,y)$ and $\gamma>0$. Thus,

$$\hat{T}_{11}=\hat{T}_{21}=0.$$

Next, fix $\mathbf{z}$ and $y>0$ arbitrarily, and let

$$\begin{aligned}\mathscr{P}&=\mathscr{P}_{\mathbf{z}},\\ \mathscr{C}_y&=\{(x,\bar{y}):|\bar{y}|<y\},\\ \ell_y&=\{(x,\bar{y}):\bar{y}=y\}.\end{aligned}$$

It then follows that

$$\begin{aligned}\mathbf{0}&=\int_{\mathscr{P}\cap\mathscr{C}_y}\operatorname{div}\mathbf{S}\,dv=\int_{\partial(\mathscr{P}\cap\mathscr{C}_y)}\mathbf{S}\mathbf{n}\,da\\ &=\int_{\mathscr{P}\cap\ell_y}\mathbf{S}\mathbf{e}^2\,da+\int_{\mathscr{S}_{-y}}\mathbf{S}\mathbf{n}\,da+\int_{\mathscr{L}_{\mathbf{z}}\cap\mathscr{C}_y}\mathbf{S}\mathbf{n}\,da+\int_{\mathscr{L}_{\mathbf{z}+\mathbf{p}}\cap\mathscr{C}_y}\mathbf{S}\mathbf{n}\,da,\end{aligned} \tag{7.15}$$

where $\mathscr{L}_{\mathbf{z}}$ and $\mathscr{L}_{\mathbf{z}+\mathbf{p}}$ are the left and right faces of $\mathscr{P}$ and $\mathscr{S}_{-y}$ is as defined by (6.1). **S**, however, is periodic and, for each **x** in $\mathscr{L}_{\mathbf{z}}$, $\mathbf{n}(\mathbf{x})=-\mathbf{n}(\mathbf{x}+\mathbf{p})$. Hence, the last two integrals in (7.15) cancel each other. What remains can be written:

$$\frac{-1}{p}\int_{\mathscr{S}_{-y}}\mathbf{Sn}\,da=\frac{1}{p}\int_{\mathscr{P}\cap\ell_y}\mathbf{Se}^2\,da. \tag{7.16}$$

The limit, as y approaches infinity, of the right-hand side of (7.16), can easily be computed using (7.13) and the definitions of **S*** and $\hat{\mathbf{S}}$. The result is

$$\lim_{y\to\infty}\frac{-1}{p}\int_{\mathscr{S}_{-y}}\mathbf{Sn}\,da=\hat{\mathbf{S}}\mathbf{e}^2 \tag{7.17}$$

proving, in a single step, that $\bar{\mathbf{s}}$ exists and that

$$\hat{S}_{i2}=\bar{s}_i$$

for $i=1$ and 2. □

By the above, it makes sense to discuss the vector $\mathbf{Se}^2$ "at $y=+\infty$". Likewise, if $\mathscr{B}$ contains a lower half plane (i.e., a set of the form $\{(x,y):y<y^0\}$ for some y^0), then $\mathbf{Se}^2$ "at $y=-\infty$" is also a well-defined fixed vector. These two tractions, however, may not be equal. Indeed, if $\mathscr{B}$ does contain a lower half plane, then (7.15) leads to

$$\frac{-1}{p}\int_{\partial\mathscr{B}\cap\partial\mathscr{P}}\mathbf{s}\,da-\mathbf{Se}^2|_{y=-\infty}=\bar{\mathbf{s}}=\mathbf{Se}^2|_{y=+\infty}$$

from which the following observations may be made:

1. If $\mathscr{B}$ is contained in a single half plane, then $\mathbf{Se}^2|_{y=+\infty}$ is determined entirely by **s** on $\partial\mathscr{B}$ through the formula

$$\mathbf{Se}^2|_{y=+\infty}=\frac{-1}{p}\int_{\partial\mathscr{B}\cap\partial\mathscr{P}}\mathbf{s}\,da.$$

2. If $\mathscr{B}$ contains both upper and lower half planes, then $\mathbf{Se}^2|_{y=+\infty}$ equals $\mathbf{Se}^2|_{y=-\infty}$ if and only if

$$\int_{\partial\mathscr{B}\cap\partial\mathscr{P}}\mathbf{s}\,da=\mathbf{0}.$$

8. Consequences. It would be unnatural not to discuss the implications of Theorem 7.1 with regard to questions concerning the finiteness of the total energy, the uniqueness of solutions, Saint-Venant's principle, etc. We shall not be unnatural. To simplify the discussion it will be assumed in this section that any body, $\mathscr{B}$, satisfies all of the following in addition to being homogeneous and isotropic.

1. $\mathscr{B}$ is two-dimensional.
2. The characteristic function for $\mathscr{B}$ is periodic (with period **p**).
3. $\partial\mathscr{B}\cap\partial\mathscr{P}$ is a bounded set in space.

The first theorem is simply the observation that Theorem 7.1 can be viewed as a sort of Saint-Venant's principle. Details of the proof will be left to the reader.

THEOREM 8.1 (Saint-Venant's principle). *Let* $(\mathbf{u}^1, \mathbf{E}^1, \mathbf{S}^1)$ *and* $(\mathbf{u}^2, \mathbf{E}^2, \mathbf{S}^2)$ *be two periodic strain solutions to two, possibly different, periodic boundary value problems on* $\mathscr{B}$ *both involving the same body force. Assume that one of the following holds*:

i. $\mathbf{u}^1(\mathbf{x}) - \mathbf{u}^2(\mathbf{x}) = o(|\mathbf{x}|)$ *as* $|\mathbf{x}| \to \infty$ *on* $\mathscr{P}$;

ii. $\mathbf{S}^1(\mathbf{x}) - \mathbf{S}^2(\mathbf{x}) = o(1)$ *as* $|\mathbf{x}| \to \infty$ *on* $\mathscr{P}$;

iii. $\mathscr{S}_1^i$, *the surface on which* $\mathbf{u}^i$ *is prescribed, is nontrivial for both* $i = 1$ *and* $i = 2$, *and either*

$$\mathbf{S}^1\mathbf{e}^1 - \mathbf{S}^2\mathbf{e}^1 = o(1) \quad \text{as } |\mathbf{x}| \to \infty \quad \text{on } \mathscr{P}$$

or

$$\mathbf{S}^1\mathbf{e}^2 - \mathbf{S}^2\mathbf{e}^2 = o(1) \quad \text{as } |\mathbf{x}| \to \infty \quad \text{on } \mathscr{P};$$

iv. $\mathscr{B}$ *is contained in a single half plane and both of the following hold*:

$$\int_{\partial\mathscr{B}\cap\partial\mathscr{P}} \mathbf{S}^1\mathbf{n}\, da = \int_{\partial\mathscr{B}\cap\partial\mathscr{P}} \mathbf{S}^2\mathbf{n}\, da,$$

$$S_{11}^1(\mathbf{x}) - S_{11}^2(\mathbf{x}) = o(1) \quad \text{as } |\mathbf{x}| \to \infty \quad \text{on } \mathscr{P}.$$

Then, for some fixed vector, $\mathbf{v}^0$, *and any positive constant,* γ,

$$\mathbf{u}^1(\mathbf{x}) - \mathbf{u}^2(\mathbf{x}) - \mathbf{v}^0 = o\left(|\mathbf{x}|^{-\gamma}\right) \quad \text{as } |\mathbf{x}| \to \infty \quad \text{on } \mathscr{P},$$

$$\mathbf{E}^1(\mathbf{x}) - \mathbf{E}^2(\mathbf{x}) = o\left(|\mathbf{x}|^{-\gamma}\right) \quad \text{as } |\mathbf{x}| \to \infty \quad \text{on } \mathscr{P},$$

$$\mathbf{S}^1(\mathbf{x}) - \mathbf{S}^2(\mathbf{x}) = o\left(|\mathbf{x}|^{-\gamma}\right) \quad \text{as } |\mathbf{x}| \to \infty \quad \text{on } \mathscr{P}.$$

The next theorem is the extension of the well-known theorem of work and energy which states that if $(\mathbf{u}, \mathbf{E}, \mathbf{S})$ is an elastic state corresponding to a body force $\mathbf{b}$ on a *bounded* body, Ω, then

$$\int_\Omega \mathbf{E}\cdot\mathbf{S}\, dv = \int_{\partial\Omega} \mathbf{u}\cdot\mathbf{s}\, da + \int_\Omega \mathbf{u}\cdot\mathbf{b}\, dv.$$

The main difficulty in extending the classical theorem of work and energy to elastic states on unbounded bodies lies in the difficulty of assuring the convergence of the resulting improper integrals without making undesirably strong assumptions on the behavior of the state "near infinity". Fortunately, Theorem 7.1 insures that periodic strain states on $\mathscr{B}$ do approach their values "at infinity" quite rapidly. Thus, in this case, extending the theorem of work and energy is a very simple exercise and will be left to the reader.

THEOREM 8.2 (work and energy). *Let* $(\mathbf{u}, \mathbf{E}, \mathbf{S})$ *be a periodic strain state corresponding to zero body force on* $\mathscr{B}$ *which satisfies either*

$$\mathbf{u}(\mathbf{x}) = o(|\mathbf{x}|) \quad \text{as } |\mathbf{x}| \to \infty \quad \text{on } \mathscr{P}$$

or

$$\mathbf{S}(\mathbf{x}) = o(1) \quad \text{as } |\mathbf{x}| \to \infty \quad \text{on } \mathscr{P}.$$

Then the total work expended in the deformation of any period section, $\mathscr{P}$, *is finite and is given by*

$$\int_{\mathscr{P}} \mathbf{E}\cdot\mathbf{S}\, dv = \int_{\partial\mathscr{P}\cap\partial\mathscr{B}} \mathbf{u}\cdot\mathbf{s}\, da.$$

If, however, $(\mathbf{u}, \mathbf{E}, \mathbf{S})$ *is a periodic strain corresponding to zero body force on* $\mathscr{B}$ *which satisfies either*

$$\mathbf{u}(\mathbf{x}) = o\left(|\mathbf{x}|^2\right) \quad as\ |\mathbf{x}| \to \infty \quad on\ \mathscr{P}$$

or

$$\mathbf{S}(\mathbf{x}) = o\left(|\mathbf{x}|\right) \quad as\ |\mathbf{x}| \to \infty \quad on\ \mathscr{P},$$

then

$$\lim_{y \to \infty} \frac{1}{\mathscr{V}_y} \int_{\mathscr{P}_y^+} \mathbf{E} \cdot \mathbf{S}\, dv$$

exists and equals

$$\frac{4\mu(\mu+\lambda)}{2\mu+\lambda}\left[\frac{\omega}{p} y + \kappa\right]^2 + \frac{1}{\mu}(\bar{s}_1)^2 + \frac{1}{2\mu+\lambda}(\bar{s}_2)^2,$$

where

$$\mathscr{P}_y^+ = \mathscr{P} \cap \{(x, \bar{y}) : 0 < \bar{y} < y\}$$

and

$$\mathscr{V}_y = \text{the area of } \mathscr{P}_y^+$$

and ω, κ, $\bar{s}_1$, *and* $\bar{s}_2$ *are the constants from the definition of* $(\mathbf{u}^0, \mathbf{E}^0, \mathbf{S}^0)$ *in Theorem* 7.1.

Later (in §9), a theorem of work and energy will be presented in which the body force is not assumed to be zero.

The reader may have already recognized that the bodies being considered here are the two-dimensional analogues of the periodic exterior bodies discussed in [4]. Indeed, the restriction that periodic exterior bodies be three-dimensional could have been dropped without affecting any of the theorems (Theorems 8.1 through 8.5 in [4]). The advantage of waiting until now to discuss these theorems with two-dimensional problems lies, of course, in the fact that, with Theorem 7.1, these same theorems can be simplified and strengthened. The resulting improvements are summarized in the two theorems below. The proofs will not be given. To simplify the statements of the theorems, the difference operator, δ, will be used where, for any suitable function, ϕ,

$$\delta\phi(\mathbf{x}) = \delta^1\phi(\mathbf{x}) = \phi(\mathbf{x}+\mathbf{p}) - \phi(\mathbf{x}),$$
$$\delta^{n+1}\phi = \delta\delta^n\phi.$$

The reader is reminded, once again, that in all theorems in this section, $\mathscr{B}$ is two-dimensional.

THEOREM 8.3 (periodicity of solutions). *Let* $\mathscr{B}$ *have positive definite elasticity and let* $(\mathbf{u}, \mathbf{E}, \mathbf{S})$ *be a solution to a periodic boundary value problem on* $\mathscr{B}$. *Assume that there are real constants,* C *and* β, *such that, for some nonnegative integer,* n, *either*

$$|\delta^n \mathbf{u}(x,y)| \leqq C\left(1 + |y|^\beta\right)\left(1 + |x|^{1/2}\right)$$

for every (x,y) *in* $\mathscr{B}$, *or*

$$|\delta^n \nabla \mathbf{E}(x,y)| \leqq C\left(1 + |y|^\beta\right)\left(1 + |x|^{1/2}\right)$$

for every (x,y) *in* $\mathscr{B}$. *Then* $\mathbf{E}$ *and* $\mathbf{S}$ *are periodic provided any one of the following hold*:

1. $\mathbf{u}(\mathbf{x}) = o(|\mathbf{x}|^2)$ *as* $|\mathbf{x}| \to \infty$ *on* $\mathscr{P}$;
2. $\mathbf{S}(\mathbf{x}) = o(|\mathbf{x}|)$ *as* $|\mathbf{x}| \to \infty$ *on* $\mathscr{P}$;
3. $\mathscr{S}_1$, *the surface on which* $\mathbf{u}$ *is prescribed, is nontrivial and either*

$$\mathbf{Se}^1 = o(|\mathbf{x}|) \quad \text{as } |\mathbf{x}| \to \infty \quad \text{on } \mathscr{P}$$

or

$$\mathbf{Se}^2 = o(|\mathbf{x}|) \quad \text{as } |\mathbf{x}| \to \infty \quad \text{on } \mathscr{P};$$

4. $\mathscr{B}$ *is contained in a single half plane, the traction is prescribed on almost all of* $\partial\mathscr{B}$, *and*

$$S_{11}(\mathbf{x}) = o(|\mathbf{x}|) \quad \text{as } |\mathbf{x}| \to \infty \quad \text{on } \mathscr{P};$$

5. *for each* $\mathbf{x}$ *in* $\mathscr{B}$

$$\delta\mathbf{S}(\mathbf{x} + k\mathbf{p}) = o(1) \quad \text{as } k \to \infty \quad \text{on } \mathbb{N};$$

6. *for each* $\mathbf{x}$ *in* $\mathscr{B}$

$$\delta\nabla\mathbf{S}(\mathbf{x} + k\mathbf{p}) = o(1) \quad \text{as } k \to \infty \quad \text{on } \mathbb{N}$$

and there are two points on $\partial\mathscr{B}$ *at which* $\mathbf{s}$ *is prescribed such that the normals to* $\partial\mathscr{B}$ *at these two points span* $\mathbb{R}^2$;

7. *for some nonnegative integer,* m, *and each* $\mathbf{x}$ *in* $\mathscr{B}$

$$\nabla^m\mathbf{u}(\mathbf{x} + k\mathbf{p}) = o(1) \quad \text{as } k \to \infty \quad \text{on } \mathbb{N}$$

and the projection of $\mathscr{S}_1$ *onto the* X_2*-axis has nontrivial interior.*

THEOREM 8.4 (uniqueness of solutions to slightly periodic problems). *Let* $\mathscr{B}$ *have positive definite elasticity and let* $(\mathbf{u}, \mathbf{E}, \mathbf{S})$ *be the difference between two solutions to the same slightly periodic boundary value problem on* $\mathscr{B}$. *Assume that there are real constants,* C *and* β, *such that, for some nonnegative integer,* n, *either*

$$|\delta^n\mathbf{u}(x,y)| \leqq C\left(1 + |y|^\beta\right)\left(1 + |x|^{1/2}\right)$$

for every (x,y) *in* $\mathscr{B}$, *or*

$$|\delta^n\nabla\mathbf{E}(x,y)| \leqq C\left(1 + |y|^\beta\right)\left(1 + |x|^{1/2}\right)$$

for every (x,y) *in* $\mathscr{B}$. *If any one of the following conditions holds, then* $\mathbf{E}$ *and* $\mathbf{S}$ *vanish throughout* $\mathscr{B}$ *and* $\mathbf{u}$ *is a rigid displacement*:

1. $\mathbf{u}(\mathbf{x}) = o(|\mathbf{x}|)$ *as* $|\mathbf{x}| \to \infty$ *on* $\mathscr{P}$;
2. $\mathbf{S}(\mathbf{x}) = o(1)$ *as* $|\mathbf{x}| \to \infty$ *on* $\mathscr{P}$;
3. $\mathscr{S}_1$, *the surface on which* $\mathbf{u}$ *is prescribed, is nontrivial and either*

$$\mathbf{Se}^1 = o(1) \quad \text{as } |\mathbf{x}| \to \infty \quad \text{on } \mathscr{P}$$

or

$$\mathbf{Se}^2 = o(1) \quad \text{as } |\mathbf{x}| \to \infty \quad \text{on } \mathscr{P};$$

4. $\mathscr{B}$ *is contained in a single half plane, the traction is prescribed on almost all of* $\partial\mathscr{B}$, *and*

$$S_{11}(\mathbf{x}) = o(1) \quad \text{as } |\mathbf{x}| \to \infty \quad \text{on } \mathscr{P};$$

5. *for each* $\mathbf{x}$ *in* $\mathscr{B}$

$$\mathbf{S}(\mathbf{x}+k\mathbf{p})=o(1) \quad \text{as } k\to\infty \quad \text{on } \mathbb{N};$$

6. *for each* $\mathbf{x}$ *in* $\mathscr{B}$

$$\nabla\mathbf{S}(\mathbf{x}+k\mathbf{p})=o(1) \quad \text{as } k\to\infty \quad \text{on } \mathbb{N};$$

and there are two points on $\partial\mathscr{B}$ *at which* $\mathbf{s}$ *is prescribed such that the normals to* $\partial\mathscr{B}$ *at these two points span* $\mathbb{R}^2$;

7. *for some nonnegative integer,* m, *and each* $\mathbf{x}$ *in* $\mathscr{B}$

$$\nabla^m\mathbf{u}(\mathbf{x}+k\mathbf{p})=o(1) \quad \text{as } k\to\infty \quad \text{on } \mathscr{B}$$

and the projection of $\mathscr{S}_1$ *onto the* X_2*-axis has nontrivial interior.*

It should be noted that conditions 5, 6, and 7 in the above two theorems were discussed in the corresponding theorems in [4]. If one of these three conditions is the condition which holds, then Theorem 7.1 need not be used in proving the claimed periodicity or uniqueness.

The final theorems of this section involve the traction problem. The restrictions on the Lamé moduli will be greatly weakened, while an additional condition (along with those assumed at the beginning of this section) will be imposed on the geometry of $\mathscr{B}$. That condition is

4. There is a fixed vector, $\mathbf{m}^0$, such that

$$\mathbf{m}^0\cdot\mathbf{n}>0$$

almost everywhere on the boundary of $\mathscr{B}$.

The term "quasi-half plane" will be used (with some apologies) to denote a body which satisfies Condition 4 as well as the conditions already assumed at the beginning of this section. The only restriction on the Lamé moduli is that

$$\mu(2\mu+\lambda)(\mu+\lambda)\neq 0.$$

The first result in this final sequence of theorems is the obvious improvement of Theorem 4.3 using Theorem 7.1. The proof is straightforward and will be omitted.

THEOREM 8.5 (uniqueness of periodic strain solutions). *Let* $\mathscr{B}$ *be a quasi-half plane and let* $(\mathbf{u},\mathbf{E},\mathbf{S})$ *be the difference between two solutions to the same traction problem on* $\mathscr{B}$. *Suppose that* $\mathbf{E}$ *is periodic and, for some* $\beta>0$, *satisfies*

$$\mathbf{E}(\mathbf{x})=O\big(|\mathbf{x}|^{\beta}\big) \quad \text{as } |\mathbf{x}|\to\infty \quad \text{on } \mathscr{P}.$$

If either

$$\mathbf{u}(\mathbf{x})=o(|\mathbf{x}|) \quad \text{as } |\mathbf{x}|\to\infty \quad \text{on } \mathscr{P}$$

or

$$S_{11}(\mathbf{x})=o(1) \quad \text{as } |\mathbf{x}|\to\infty \quad \text{on } \mathscr{P}$$

then $\mathbf{E}$ *and* $\mathbf{S}$ *vanish throughout* $\mathscr{B}$ *and* $\mathbf{u}$ *is a rigid displacement.*

In [4] it was shown that combining a general uniqueness theorem with a corresponding uniqueness theorem for periodic strain states and successive applications of Theorem 4.2 led to the assertion that certain periodic boundary value problems had only periodic strain solutions (see Theorems 8.1 and 8.3 of [4]). Likewise, an intelligent combination of Theorems 2.2 and 8.5, above, along with several applications of Theorem 3.1 will prove the following theorem. The proof however, will be omitted since it is both tedious and similar in spirit to those in [4].

THEOREM 8.6 (periodicity of solutions). *Let $\mathscr{B}$ be a quasi-half plane and let* $(\mathbf{u}, \mathbf{E}, \mathbf{S})$ *be a solution to a periodic traction problem on $\mathscr{B}$. Suppose that for some $\beta > 0$,*

$$\nabla \mathbf{u}(\mathbf{x}+\mathbf{p}) - \nabla \mathbf{u}(\mathbf{x}) = O\left(|y|^{\beta}\right) \quad \text{as } |y| \to \infty$$

and

$$\frac{\mathbf{S}(\mathbf{x}+\mathbf{p}) - \mathbf{S}(\mathbf{x})}{1+|y|^{\beta}} = o(1) \quad \text{as } |x| \to \infty$$

where $(x,y) = \mathbf{x}$. Then $\mathbf{E}$ *and* $\mathbf{S}$ *are periodic.*

Finally, consider an arbitrary, not necessarily periodic, traction problem on a quasi-half plane. By the definition of quasi-half plane the corresponding null traction problem is periodic. Thus, Theorem 8.6 can be applied using the difference between any two solutions. Once the difference is found to be periodic, Theorem 8.5 can be invoked to prove that, in fact, the difference is a trivial elastic state. This analysis is summarized in the following theorem.

THEOREM 8.7 (uniqueness in the general traction problem). *Let $\mathscr{B}$ be a quasi-half plane, and let* $(\mathbf{u}, \mathbf{E}, \mathbf{S})$ *be the difference between two solutions to the same traction problem on $\mathscr{B}$. Suppose that for some $\beta > 0$,*

$$\mathbf{E}(\mathbf{x}) = O\left(|\mathbf{x}|^{\beta}\right) \quad \text{as } |\mathbf{x}| \to \infty \quad \text{on } \mathscr{P},$$

$$\nabla \mathbf{u}(\mathbf{x}+\mathbf{p}) - \nabla \mathbf{u}(\mathbf{x}) = O\left(|y|^{\beta}\right) \quad \text{as } |y| \to \infty$$

and

$$\frac{\mathbf{S}(\mathbf{x}+\mathbf{p}) - \mathbf{S}(\mathbf{x})}{1+|y|^{\beta}} = o(1) \quad \text{as } |x| \to \infty$$

where $(x,y) = \mathbf{x}$. If either

$$\mathbf{u}(\mathbf{x}) = o(|\mathbf{x}|) \quad \text{as } |\mathbf{x}| \to \infty \quad \text{on } \mathscr{P}$$

or

$$S_{11}(\mathbf{x}) = o(1) \quad \text{as } |\mathbf{x}| \to \infty \quad \text{on } \mathscr{P}$$

or

$$S_{11}(\mathbf{x}) = o(1) \quad \text{as } |x| \to \infty,$$

then $\mathbf{E}$ *and* $\mathbf{S}$ *vanish throughout $\mathscr{B}$ and* $\mathbf{u}$ *is a rigid displacement.*

9. Nonvanishing body forces. Although the major lemma, Lemma 6.1, was derived without regard to the presence or absence of a nontrivial body force field, $\mathbf{b}$, the vanishing of $\mathbf{b}$ was an assumption in Theorem 7.1. If, instead, one assumes in Theorem 7.1 that the body force field, $\mathbf{b}$, is curl-free and divergence-free and satisfies

$$\mathbf{b}(x,y) = O(y^{-\alpha}) \quad \text{as } y \to \infty$$

for some real constant, α, then (7.9) in the proof of Theorem 7.1 becomes

$$\nabla \nabla \mathbf{u}(x,y) = O(y^{-\alpha}) \quad \text{as } y \to \infty.$$

If, in addition

$$\lim_{y\to\infty}\int_{\mathscr{P}\cap\mathscr{C}_y}\mathbf{b}\,da=0$$

where

$$\mathscr{C}_y=\{(x,\bar{y}):\bar{y}<y\},$$

then slight modifications in (7.15) and (7.16) still imply (7.17). With these changes, Theorem 7.1 becomes

THEOREM 9.1. *Suppose that $\mathscr{B}$ is a two-dimensional elastic body containing the upper half plane, $\{(x,y):y>0\}$, and that $(\mathbf{u},\mathbf{E},\mathbf{S})$ is a periodic strain state on $\mathscr{B}$ corresponding to a body force, $\mathbf{b}$, satisfying both*

$$\lim_{y\to\infty}\int_{\mathscr{P}\cap\mathscr{C}_y}\mathbf{b}\,da=0$$

and, for some $\alpha>0$,

$$\mathbf{b}(x,y)=O(y^{-\alpha})\quad \text{as } y\to\infty.$$

Assume, also, that for some arbitrary real β and some period section $\mathscr{P}$, either

$$\mathbf{u}(\mathbf{x})=O(|\mathbf{x}|^{\beta})\quad \text{as } |\mathbf{x}|\to\infty\quad \text{on } \mathscr{P}$$

or

$$\mathbf{E}(\mathbf{x})=O(|\mathbf{x}|^{\beta})\quad \text{as } |\mathbf{x}|\to\infty\quad \text{on } \mathscr{P}.$$

If $\alpha>1$, then the negative mean surface traction, $\bar{\mathbf{s}}$, exists and

$$\mathbf{E}(x,y)-\mathbf{E}^0(x,y)=O(y^{1-\alpha})\quad \text{as } y\to\infty,$$
$$\mathbf{S}(x,y)-\mathbf{S}^0(x,y)=O(y^{1-\alpha})\quad \text{as } y\to\infty$$

where $(\mathbf{u}^0,\mathbf{E}^0,\mathbf{S}^0)$ is the elastic state defined in §7. Moreover, if $\alpha>2$, then there is a fixed vector, $\mathbf{v}^0$, such that

$$\mathbf{u}(x,y)-\mathbf{u}^0(x,y)-\mathbf{v}^0=O(y^{2-\alpha})\quad \text{as } y\to\infty.$$

From this can be derived the following theorem of work and energy (compare with Theorem 8.2). $\mathscr{B}$ is as described in §8.

THEOREM 9.2 (work and energy). *Let $(\mathbf{u},\mathbf{E},\mathbf{S})$ be a periodic strain state on $\mathscr{B}$ corresponding to a curl free and divergence free body force, $\mathbf{b}$. Assume that*

$$\lim_{y\to\infty}\int_{\mathscr{P}\cap\mathscr{C}_y}\mathbf{b}\,dv=0$$

and that, for some $\alpha>2$,

$$\mathbf{b}(\mathbf{x})=O(|\mathbf{x}|^{-\alpha})\quad \text{as } |\mathbf{x}|\to\infty\quad \text{on } \mathscr{P}$$

and that either

$$\mathbf{u}(\mathbf{x})=o(|\mathbf{x}|)\quad \text{as } |\mathbf{x}|\to\infty\quad \text{on } \mathscr{P}$$

or

$$\mathbf{S}(\mathbf{x}) = o(1) \quad as\ |\mathbf{x}| \to \infty \quad on\ \mathscr{P}.$$

Then the total work expended in the deformation of any period section, $\mathscr{P}$, is finite and is given by

$$\int_{\mathscr{P}} \mathbf{E} \cdot \mathbf{S}\, dv = \int_{\partial\mathscr{B} \cap \partial\mathscr{P}} \mathbf{u} \cdot \mathbf{s}\, da + \int_{\mathscr{P}} \mathbf{u} \cdot \mathbf{b}\, dv.$$

In a similar fashion, versions of Saint-Venant's principle (Theorem 8.1) involving nonvanishing body forces can be derived as direct consequences of Theorem 9.1. These will be left to the interested reader.

10. Asymptotic behavior of periodic nonplanar strain states. There does not exist a theorem analogous to Theorem 7.1 for nonplanar strain states. This can be seen by taking the two base periodic states from §4 associated with an arbitrary three-dimensional periodic strain state in which $\omega_1 = \omega_2 = 0$. Letting $\mathscr{B}$ be the upper half space, $\{(x,y,z) : z > 0\}$, it is clear that the negative mean strain, $\bar{\mathbf{s}}$, associated with each base state is $\mathbf{0}$. Thus, provided a theorem analogous to 7.2 did hold for such states, the two base states should approach the same "asymptotic state" as z approaches infinity. Clearly, however, they do not.

REFERENCES

[1] M. E. GURTIN AND E. STERNBERG, *Theorems in linear elastostatics for exterior domains*, Arch. Rational Mech. Anal., 8 (1961), pp. 99–119.

[2] K. B. HOWELL, *Uniqueness in linear elastostatics for problems involving unbounded bodies*, J. Elasticity, 10 (1980), pp. 407–427.

[3] _____, *Periodic and "slightly" periodic boundary value problems in elastostatics on bodies unbounded in several directions*, Int. J. Engng. Sci., 20 (1982), pp. 455–481.

[4] _____, *Directionally dependent asymptotic behavior of biharmonic functions with applications to elasticity*, this Journal, 16 (1985), pp. 822–847.

[5] R. J. KNOPS AND L. E. PAYNE, *Uniqueness Theorems in Linear Elasticity*, Springer-Verlag, Berlin, 1971.

[6] N. I. MUSKHELISHVILI, *Some Basic Problems of the Mathematical Theory of Elasticity*, Trans., J. R. M. Radok, Noordhoff, Groningen, 1953.

SIAM J. MATH. ANAL.
Vol. 17, No. 1, January 1986

QUASI INNER PRODUCTS OF ANALYTIC FUNCTIONS WITH APPLICATIONS TO SPECIAL FUNCTIONS*

L. R. BRAGG†

Abstract. Let $f(z)=\sum_{n=0}^{\infty}a_n z^n$, $g(z)=\sum_{n=0}^{\infty}b_n z^n$ be analytic functions in a disk D. For $z\in D$, define the quasi inner product of f and g by

$$f(z)\circ g(z)=\frac{1}{2\pi}\cdot\int_0^{2\pi} f(ze^{i\theta})g(ze^{-i\theta})\,d\theta=\sum_{n=0}^{\infty}a_n b_n z^{2n}.$$

In this paper, (i) we treat the analytic and algebraic properties of $\circ$ and (ii) apply this composition to special functions. Included are treatments of the hypergeometric functions, generating functions, and the Lerch transcendent function. Certain generalizations of the composition $\circ$ are also considered.

AMS-MOS subject classifications (1980). Primary 33A35; secondary 30D10

Key words. analytic function, quasi inner products, diagonal of product series, special functions, generating functions, representations

1. Introduction. Let x,y,z denote complex variables and let $f(z)$, $g(z)$ be a pair of functions of z that are analytic in a disk D of radius R with center 0. Suppose that $f(z)=\sum_{n=0}^{\infty}a_n z^n$ and $g(z)=\sum_{n=0}^{\infty}b_n z^n$. Then if x, $y\in D$, we define the compositions $f(x)\circ g(y)$ of f and g by the relation

$$f(x)\circ g(y)=\frac{1}{2\pi}\cdot\int_0^{2\pi} f(xe^{i\theta})g(ye^{-i\theta})\,d\theta=\sum_{n=0}^{\infty}a_n b_n x^n y^n. \tag{1.1}$$

The last member of this is simply the sum of the main diagonal terms in the product of the series for $f(x)$ and $g(y)$. The integral in (1.1) defines a convolution of two functions and is a modified version of a contour integral employed by M. L. J. Hautus and D. A. Klarner in their treatment of the diagonal terms of a double power series whose coefficients are defined by a linear recurrence relation [8]. In addition, J. Hadamard made use of an integral analogous to (1.1) to determine the singularities of the function defined by the series $\sum_{n=0}^{\infty}a_n b_n z^n$ (Hadamard's multiplication theorem) [13]. If the a_n and b_n are real, then the right member of (1.1) assigns a real value to each fixed pair of reals x, $y\in D$ and has the form of an inner product. Similar inner products appear in H^2 space theory [5], [9], [12]. In our discussions, we permit x and y to vary in D in (1.1) and refer to the composition in (1.1) as a *quasi inner product* which we abbreviate by qip.

Our interest in qip (1.1) (and a number of its generalizations) is motivated by (i) its utility in developing integral representations and related results for special functions and (ii) its applicability to transmutations (see [1], [3]) and the representation of solutions of partial differential equations. Qip's can be used as function builders and (1.1) permits the introduction of multipliers. Thus, if $g(y)$ is analytic in a region which includes $|y|=1$, then by selecting $y=1$ in (11.) we have the formula

$$\sum_{n=0}^{\infty}a_n b_n x^n=\frac{1}{2\pi}\cdot\int_0^{2\pi} f(xe^{i\theta})g(e^{-i\theta})\,d\theta \tag{1.2}$$

*Received by the editors February 3, 1983, and in revised form March 14, 1984.
†Department of Mathematical Sciences, Oakland University, Rochester, Michigan 48063.

in which the coefficients a_n in the expansion of $f(x)$ are multiplied by the b_n's (see [6]). As an example of (i) suppose we select $f(x)=g(z)=e^{z/2}$. Then (1.1) immediately yields the familiar result:

$$\frac{1}{2\pi}\cdot\int_0^{2\pi} e^{z\cos\theta}\,d\theta = e^{z/2}\circ e^{z/2} = \sum_{n=0}^{\infty}\frac{z^{2n}}{2^{2n}(n!)^2} = I_0(z),$$

a modified Bessel function of index 0. By making other suitable choices for $f(z)$, $g(z)$ in (1.1) as special functions or generators of special functions, one can easily obtain numerous integration results that would, at best, be more tedious using standard methods. With modifications in the right member of (1.1), one can also construct superdiagonal sums (i.e. $\Sigma_{n=0}^{\infty}a_nb_{n+j}x^ny^{n+j}j\geqq 1$) or subdiagonal sums ($\Sigma_{n=0}^{\infty}a_{n+j}b_nx^{n+j}y^n, j\geqq 1$) from the product $f(x)\cdot g(y)$. For instance, if we replace $f(x)$ in (1.1) by $x^jf(x), j$ a positive integer, it is easy to verify that

$$\frac{1}{2\pi}\cdot\int_0^{2\pi} e^{ji\theta}f(xe^{i\theta})g(ye^{-i\theta})\,d\theta = \sum_{n=0}^{\infty}a_nb_{n+j}x^ny^{n+j}. \tag{1.3}$$

Again, taking $f(z)=g(z)=e^{z/2}$ in this, it follows that

$$I_j(z)=\frac{1}{2\pi}\cdot\int_0^{2\pi} e^{ji\theta}e^{z\cos\theta}\,d\theta=\frac{1}{2\pi}\cdot\int_0^{2\pi}(\cos j\theta)\,e^{z\cos\theta}\,d\theta.$$

One can construct other qip's such as

$$f(z)\,{}_p\circ_q\, g(y)=\frac{1}{2\pi}\cdot\int_0^{2\pi} f(xe^{pi\theta})g(ye^{-qi\theta})\,d\theta \tag{1.4}$$

with p,q positive integers and, for simplicity, $(p,q)=1$. This will select out sums of terms from $f(x)\cdot g(y)$ along rays other than the usual diagonals. Modifications similar to (1.3) can also be developed for this qip. The qip's (1.1) and (1.4) bring together notions from both analytic functions and Fourier analysis.

In this paper, we will treat the basic properties of these qip's and certain of their applications to special functions while deferring their uses in p.d.e's. Section 2 will be concerned with the analytic properties of qip's (1.1.) and (1.4). By making use of relations such as (1.3) for superdiagonal and subdiagonal sums, we construct a partial sum integral that approximates the product $f(x)\cdot g(y)$. Not surprising, the kernel function in this approximation is the same as the one in the Fourier series case and leads to analogous convergence results. We also note the algebraic and differentiation properties of these quasi inner products. Finally, in §§3–5, we apply these qip's to obtain a variety of representations and results for special functions. Many of these appear to be new while, in other cases, the method leads to simplifications over standard treatments. In §3, we apply these techniques to the hypergeometric functions ${}_pF_q$. We apply (1.1) and (1.4) in §4 to a number of generating functions of special polynomials and functions. In §5, we note some results for the Lerch transcendent function. Finally, we consider some examples that involve extensions of the ideas behind the compositions $\circ$ and ${}_p\circ_q$.

2. Analytic properties of (1.1) and (1.4). If $f(x)$, $g(y)$ are analytic for x, $y\in D$, then it readily follows that $f(x)\circ g(y)$ is analytic in x and y. For, there exists $\rho>0$ with $\max(|x|,|y|)<\rho<R$ and $M>0$, $N>0$ such that $|a_n|<M/\rho^n$, $|b_n|<N/\rho^n$. Hence

$$\left|\frac{1}{2\pi}\cdot\int_0^{2\pi} f(xe^{i\theta})g(ye^{-i\theta})\,d\theta\right|\leqq\sum_{n=0}^{\infty}|a_n||b_n||xy|^n\leqq\sum_{n=0}^{\infty}MN\frac{|xy|^n}{\rho^{2n}}$$

and this last series converges by the ratio test. Thus, $f(x) \circ g(y)$ is well defined for x, $y \in D$.

If at least one of the two functions appearing in the qip (1.1) is entire, say $g(z)$, then one can prove the following result.

THEOREM 2.1. *Let $f(x)$ be analytic in x with*

$$\limsup_{n \to \infty} |a_n|^{1/n} = \frac{1}{R}, \qquad 0 < R < \infty,$$

and let $g(y)$ be entire in y. Then for $|x| < R$, $f(x) \circ g(y)$ is entire in xy.

In the proof of Theorem 2.1, the restriction imposed on x is necessitated by requiring the integral in (1.1) to be well defined. However, the right member of (1.1) is well defined for all xy. In practice, one can select $x \in D$, $x \neq 0$, and then choose y so that xy takes on the value one wishes to use.

In addition to these results, there are a variety of algebraic and other analytic properties associated with the qip (1.1). The following can be easily verified by the reader.

THEOREM 2.2. *Let $f(z)$, $g(z)$, and $h(z)$ be analytic for $z \in D$. Then for $x, y, z \in D$, we have*

(i) $f(x) \circ g(y) = g(y) \circ f(x)$;

(ii) $f(x) \circ [g(y) + h(y)] = f(x) \circ g(y) + f(x) \circ h(y)$;

(iii) *if $f(z)$ and $g(z)$ have no powers of z in common in their expansions, then* $f(z) \circ g(z) = 0$;

(iv) *for n a nonnegative integer.*

$$z^n \frac{d^n}{dz^n}(f(z) \circ g(z)) = \sum_{j=0}^{n} \binom{n}{j} \left[\left(z^{n-j} f^{(n-j)}(z) \right) \circ \left(z^j g^{(j)}(z) \right) \right].$$

The last of these can be checked by induction. On the other hand, we have, in general,

$$f(z) \circ [g(z) \circ h(z)] \neq [f(z) \circ g(z)] \circ h(z) \tag{2.1}$$

(for example, select $f(z) = g(z) = z$ and $h(z) = z^2$). Thus, relative to the composition $\circ$, the set of functions analytic in D have divisors of 0 and are nonassociative. Further, if $g(z) = z$, then $1 \circ g(z) = 0$ so that 1 does not serve as a multiplicative identity relative to the $\circ$ composition. In using the qip (1.1) to construct "advanced" functions from elementary ones, the failure of associativity usually presents no serious problems. One computes, say, $f(z) \circ g(z)$ and then replaces z^2 in this by a new variable before making further applications of the $\circ$ composition.

In (1.3), we obtained an integral formula for the sums of terms of $f(x) \cdot g(y)$ taken from the superdiagonals. Similarly, if we replace $g(y)$ in (1.1) by $y^j \cdot g(y)$, one can verify that

$$\frac{1}{2\pi} \cdot \int_0^{2\pi} e^{-ji\theta} f(xe^{i\theta}) g(ye^{-i\theta}) \, d\theta = \sum_{n=0}^{\infty} a_{n+j} b_n x^{n+j} y^n, \qquad j \geqq 1. \tag{2.2}$$

The right member of this yields a sum of terms from $f(x) \cdot g(y)$ taken over a subdiagonal.

Finally, suppose we sum the left members of (1.3) and (2.2) on j from $-N$ to N by using (2.2) for j negative and (1.3) for j nonnegative. We get

(2.3)

$$\frac{1}{2\pi}\cdot\int_0^{2\pi} f(xe^{i\theta})g(ye^{-i\theta})\left(\sum_{j=-N}^{N} e^{ji\theta}\right)d\theta=\frac{1}{2\pi}\int_0^{2\pi} f(xe^{i\theta})g(ye^{-i\theta})\left\{\frac{\sin(N+\frac{1}{2})\theta}{\sin\frac{1}{2}\theta}\right\}d\theta.$$

Since $f(xe^{i\theta})$ and $g(ye^{-i\theta})$ are analytic and periodic in θ of period 2π for fixed x and y, it follows, from the Dirichlet convergence criterion for Fourier series ([4], [13]), that

$$\lim_{N\to\infty}\frac{1}{2\pi}\cdot\int_0^{2\pi} f(xe^{i\theta})g(ye^{-i\theta})\frac{\sin(N+\frac{1}{2})\theta}{\sin\frac{1}{2}\theta}\,d\theta=f(x)\cdot g(y). \tag{2.4}$$

Alternatively, one could assume that the sum of the diagonal, superdiagonal, and subdiagonal terms of $f(x)\cdot g(y)$ converges to $f(x)\cdot g(y)$. From this, we could then infer the limit relation (2.4).

We observe that if $f(z)$ is analytic in a disk D of radius $R>1$, then

$$\frac{1}{2\pi}\cdot\int_0^{2\pi}|f(e^{i\theta})|^2\,d\theta=\sum_{n=0}^{\infty}a_n^2,\qquad a_n \text{ real}. \tag{2.5}$$

If, for example, we make the choice

$$f(z)=\frac{1}{a-bz}=\sum_{n=0}^{\infty}\frac{b^n z^n}{a^{n+1}},\qquad a>b>o,$$

then (2.5) leads to the integral evaluation

$$\int_0^{2\pi}[a^2+b^2-2ab\cos\theta]^{-1}\,d\theta=\frac{2\pi}{a^2-b^2}.$$

Hence, if $c>d>0$, we get

$$\int_0^{2\pi}(c-d\cos\theta)^{-1}\,d\theta=\frac{2\pi}{(c^2-d^2)^{1/2}}. \tag{2.6}$$

For the qip (1.4), it is easy to show that

$$f(x)\,{}_p\circ_q\, g(y)=\sum_{n=0}^{\infty}a_{nq}b_{np}x^{nq}y^{np} \tag{2.7}$$

and the right member of this is analytic if $x, y\in D$. The properties (ii)–(iv) of Theorem 2.2 as well as the nonassociativity (2.1) are applicable to the qip ${}_p\circ_q$. We note that the commutative law fails for this qip but we do have the relation

$$f(x)\,{}_p\circ_q\, g(y)=g(y)\,{}_q\circ_p\, f(x).$$

One also obtains somewhat more general results than ones such as (1.3) or (2.6) for (1.1). In fact, a straightforward calculation shows that

$$\frac{1}{2\pi}\cdot\int_0^{2\pi} e^{ki\theta}f(xe^{pi\theta})g(ye^{-qi\theta})\,d\theta=\sum_{\substack{m,n\geqq 0\\ mq-np=k}}a_n b_m x^n y^m,\qquad k\geqq 1. \tag{2.8}$$

Suppose that m_0, n_0 are the least nonnegative integer values of m,n such that $mq-np=k$. Then we take $m=lp+m_0$, $n=lq+n_0$, $l=0,1,2,\cdots$ for this Diophantine equation. The relation (2.8) becomes

$$\frac{1}{2\pi}\cdot\int_0^{2\pi} e^{ki\theta}f(xe^{pi\theta})g(ye^{-qi\theta})\,d\theta=\sum_{l=0}^{\infty}a_{lq+n_0}b_{lp+m_0}x^{lq+n_0}y^{lp+m_0}. \tag{2.9}$$

If we desire a formula analogous to (1.3), select $m_0=j$ and $n_0=0$ in (2.9); then k is required to have the value $q\cdot m_0$ in the left member of (2.9).

3. The hypergeometric functions. Let α and β denote a pair of vectors with $\alpha=(\alpha_1,\cdots,\alpha_{p_1})$, $\beta=(\beta_1,\cdots,\beta_{q_1})$ where the α_i and β_j are real, $q_1+1\geqq p_1$, and $\beta_j>0$ for $j=1,\cdots,q_1$. Let $(\alpha)_n=\prod_{i=1}^{p_1}(\alpha_i)_n$ where $(\alpha_i)_n=\alpha_i(\alpha_i+1)\cdots(\alpha_i+n-1)$. Then we can express the generalized hypergeometric function ${}_{p_1}F_{q_1}(\alpha_1,\cdots,\alpha_{p_1};\beta_1,\cdots,\beta_{q_1};z)$ in the form

$${}_{p_1}F_{q_1}(\alpha;\beta;z)=\sum_{n=0}^{\infty}\frac{(\alpha)_n}{(\beta)_n}\frac{z^n}{n!}. \tag{3.1}$$

Next, suppose γ and δ are a second pair of vectors analogous to α and β having p_2 and q_2 components, respectively, with $q_2+1\geqq p_2$. Denote the concatenated vector $(\alpha_1,\cdots,\alpha_{p_1},\gamma_1,\cdots,\gamma_{p_2})$ of length p_1+p_2 by $\alpha\oplus\gamma$. It follows from (3.1) and (1.1) that if $|x|<1$, $|y|<1$, then

$$\begin{aligned}{}_{p_1+p_2}F_{q_1+q_2+1}(\alpha\oplus\gamma;\beta\oplus\delta\oplus 1;xy)&\\ ={}_{p_1}F_{q_1}(\alpha;\beta;x)\circ{}_{p_2}F_{q_2}(\gamma;\delta;y)&\\ =\frac{1}{2\pi}\cdot\int_0^{2\pi}{}_{p_1}F_{q_1}(\alpha;\beta;xe^{i\theta})\,{}_{p_2}F_{q_2}(\gamma;\delta;ye^{-i\theta})\,d\theta.&\end{aligned} \tag{3.2}$$

In the following, we develop integrals and related results for some special cases of (3.2). To avoid lengthy details, we will not reduce all integrands obtained in real form. In the last example, we will make use of the qip ${}_1\circ{}_2$. For our purposes, we require the following:

$$\begin{aligned}&\text{(a)}\quad {}_0F_0(-;-;z)=e^z,\\ &\text{(b)}\quad {}_1F_0(\alpha;-;z)=(1-z)^{-\alpha}=\sum_{n=0}^{\infty}\frac{(\alpha)_n}{n!}z^n,\\ &\text{(c)}\quad {}_1F_1(1;c;z)=\sum_{n=0}^{\infty}\frac{z^n}{(c)_n},\qquad c>0.\end{aligned} \tag{3.3}$$

I. *The standard Bessel functions.* In the introduction, we indicated how to use (1.3). To obtain analogous but slightly more general forms for the J_ν's, ν a nonnegative integer, suppose we select $j=\nu$, $f(x)=e^{\lambda^2 x}(\lambda>0)$, and $g(x)=e^{-x}$ in (1.3). Then that formula yields

$$\begin{aligned}\frac{1}{2\pi}\cdot\int_0^{2\pi}e^{\nu i\theta}e^{\lambda^2 xe^{i\theta}}e^{-xe^{-i\theta}}\,d\theta&=\frac{1}{2\pi}\cdot\int_0^{2\pi}e^{(\lambda^2-1)x\cos\theta}\cos\left[(\nu\theta+(\lambda^2+1)x\sin\theta)\right]d\theta\\ &=(-1)^\nu\lambda^{-\nu}J_\nu(2\lambda x)\end{aligned}$$

or

$$J_\nu(2\lambda x)=\frac{(-1)^\nu\lambda^\nu}{2\pi}\int_0^{2\pi} e^{(\lambda^2-1)x\cos\theta}\cos\left[\nu\theta+(\lambda^2+1)x\sin\theta\right]d\theta. \tag{3.4}$$

It is useful to compare the above derivation of the integral for $J_\nu(2\lambda x)$ with the standard one starting with the generating function for the Bessel functions.

II. *A* ${}_2F_1$ *function.* Suppose we select $f(z)=(1-z)^{-\alpha}$, $g(z)=(1-z)^{-\beta}$, $\alpha>0$, $\beta>0$. Then for $|z|<1$, we have, by (3.3b)

$$f(z)\circ g(z)=\frac{1}{2\pi}\cdot\int_0^{2\pi}(1-ze^{i\theta})^{-\alpha}(1-ze^{-i\theta})^{-\beta}d\theta \tag{3.5}$$

$$={}_2F_1(\alpha,\beta;1;z^2).$$

The integral in this can be expressed in the form

$$\frac{1}{2\pi}\cdot\int_0^{2\pi}\frac{\cos(\beta-\alpha)\phi}{[1-2z\cos\theta+z^2]^{(\alpha+\beta)/2}}d\theta,$$

with

$$\phi=\tan^{-1}\left(\frac{z\sin\theta}{1-z\cos\theta}\right)\quad\text{if } z \text{ is real.}$$

III. *A definite integral involving ultraspherical polynomials.* Let $\alpha=\beta$ in (3.5). Then, for $|z|<1$,

$$(1-ze^{i\theta})^{-\alpha}(1-ze^{-i\theta})^{-\alpha}=(1-2z\cos\theta+z^2)^{-\alpha},$$

this last being a generating function for the ultraspherical polynomials $P_n^\alpha(\cos\theta)$ [10]. Hence

$$f(z)\circ g(z)=\frac{1}{2\pi}\cdot\int_0^{2\pi}(1-2z\cos\theta+z^2)^{-\alpha}d\theta$$

$$=\frac{1}{2\pi}\cdot\int_0^{2\pi}\left(\sum_{n=0}^{\infty}P_n^\alpha(\cos\theta)z^n\right)d\theta \tag{3.6}$$

$$=\sum_{n=0}^{\infty}z^n\left(\frac{1}{2\pi}\cdot\int_0^{2\pi}P_n^\alpha(\cos\theta)\,d\theta\right).$$

But from (3.5) with $\beta=\alpha$, we have

$$f(z)\circ g(z)={}_2F_1(\alpha,\alpha;1;z^2)=\sum_{n=0}^{\infty}\frac{[(\alpha)_n]^2}{(n!)^2}z^{2n}. \tag{3.7}$$

Comparing the right member of (3.6) with the right member of (3.7), we obtain the familiar integral evaluation ([10])

$$\frac{1}{2\pi}\cdot\int_0^{2\pi}P_n^\alpha(\cos\theta)\,d\theta=\begin{cases}0, & n \text{ odd},\\ \left(\dfrac{(\alpha)_m}{m!}\right)^2, & n=2m.\end{cases} \tag{3.8}$$

The choice $\alpha=\frac{1}{2}$ in (3.8) gives the result for Legendre polynomials. We note the economy in obtaining (3.8) using $\circ$ as compared with standard techniques.

IV. *A* ${}_1F_1$ *type transformation.* From (3.3b, c), the reader can verify that

$$ {}_1F_1(1;c;x)\circ {}_1F_0(\alpha;-;y)={}_1F_1(\alpha;c;xy). \tag{3.9}$$

Now ${}_1F_1(1;c;x)$ is entire and ${}_1F_0(\alpha;-;y)$ is analytic with $R=1$. Theorem 2.1 shows that ${}_1F_1(\alpha;c;xy)$ is entire in xy if $|y|<1$.

V. *A* ${}_p\circ_q$ *example.* If we select $f(x)=e^x$ and $g(y)=(1-y)^{-\alpha}$, then it follows by (2.7) that

$$\begin{aligned} f(x)\,{}_1\circ_2\, g(y) &= \sum_{n=0}^{\infty} \frac{1}{(2n)!}\,\frac{(\alpha)_n}{n!}x^{2n}y^n \\ &= {}_1F_2\left(\alpha;\frac{1}{2},1;\frac{x^2y}{4}\right) \\ &= \frac{1}{2\pi}\cdot\int_0^{2\pi} e^{xe^{i\theta}}\left(1-ye^{-2i\theta}\right)^{-\alpha}d\theta. \end{aligned}$$

4. Generating functions. Quasi inner products, when used in conjunction with generating functions, are convenient for developing results for special functions. This is particularly true when the generating function is of exponential type and the Euler relations can be called upon to express integrals in a real form. In the following, we apply (1.1) and (1.4) to various generating functions to derive integral representations for certain infinite sums of products of special polynomials. We include cases that involve the Hermite, the generalized Hermite, and the ultraspherical polynomials. One example is given, without detailed development, that involves products of two different types of polynomials. Since these generating functions contain both variables and parameters, it will be necessary to distinguish which variable or parameter multiplies $e^{i\theta}$ or $e^{-i\theta}$ (respectively, $e^{pi\theta}$ or $e^{-qi\theta}$) in the choices for f and g in (1.1) (or 1.4). We do this by underscoring the designated variable or parameter.

A. *Hermite polynomials.* The Hermite polynomials are generated by means of the relation

$$f(x,t)=e^{2xt-t^2}=\sum_{n=0}^{\infty}\frac{H_n(x)t^n}{n!}. \tag{4.1}$$

If we take $g(x,t)=f(x,t)$, then

$$\begin{aligned} f(x,\underline{t})\circ f(x,\underline{t}) &= \sum_{n=0}^{\infty}\frac{[H_n(x)]^2}{(n!)^2}t^{2n} \\ &= \frac{1}{2\pi}\cdot\int_0^{2\pi}\exp(2txe^{i\theta}-t^2e^{2i\theta})\exp(2xte^{-i\theta}-t^2e^{-2i\theta})\,d\theta \\ &= \frac{1}{2\pi}\cdot\int_0^{2\pi}\exp(4tx\cos\theta-2t^2\cos 2\theta)\,d\theta. \end{aligned} \tag{4.2}$$

Similarly, the quasi inner product $f(x,\underline{t})\circ f(y,\underline{t})$ leads to the relationship

$$\begin{aligned} &\sum_{n=0}^{\infty}\frac{H_n(x)H_n(y)}{(n!)^2}t^{2n} \\ &\quad=\frac{1}{2\pi}\cdot\int_0^{2\pi}\exp(2t(x+y)\cos\theta-2t^2\cos 2\theta)\cos[2t(x-y)\sin\theta]\,d\theta. \end{aligned} \tag{4.3}$$

We note that (4.3) reduces to (4.2) when $y=x$.

B. *The generalized Hermite polynomials.* The generalized Hermite polynomials $g_n^p(x)$ are generated by the relation

$$f(x,t)=\exp(pxt-t^p)=\sum_{n=0}^{\infty}\frac{1}{n!}g_n^p(x)t^n \tag{4.4}$$

(see [7], also [2]). Using a calculation as in A above, we get

$$\begin{aligned} f(x,\underline{t})\circ f(x,\underline{t}) &= \frac{1}{2\pi}\cdot\int_0^{2\pi}\exp(2pxt\cos\theta-2t^p\cos p\theta)\,d\theta \\ &= \sum_{n=0}^{\infty}\frac{\{g_n^p(x)\}^2}{(n!)^2}t^{2n}. \end{aligned} \tag{4.5}$$

One can similarly obtain a formula analogous to (4.3).

C. *Ultraspherical polynomials.* We refered to these earlier and have, in fact

$$f(x,t)=(1-2xt+t^2)^{-\lambda}=\sum_{n=0}^{\infty}P_n^{\lambda}(x)t^n \tag{4.6}$$

with $|t|<1$ and $\lambda>0$. Forming $f(x,\underline{t})\circ f(x,\underline{t})$, we get

$$\sum_{n=0}^{\infty}\left[P_n^{\lambda}(x)\right]^2t^{2n}=\frac{1}{2\pi}\cdot\int_0^{2\pi}\left[1+4x^2t^2-4xt(1+t^2)\cos\theta+2t^2\cos 2\theta\right]^{-\lambda}d\theta. \tag{4.7}$$

D. *Hermite polynomials using* ${}_1\circ_2$. Suppose we again select $f(x,t)$ as in A above. Then

$$\begin{aligned} f(x,\underline{t})\,{}_1\circ_2 f(x,\underline{t}) &= \sum_{n=0}^{\infty}\frac{H_n(x)H_{2n}(x)}{n!(2n)!}t^{3n} \\ &= \frac{1}{2\pi}\cdot\int_0^{2\pi}\exp(2xt\cos\theta+(2xt-t^2)\cos 2\theta-t^2\cos 4\theta) \\ &\qquad\cdot\cos\left[2xt\sin\theta-(t^2+2xt)\sin 2\theta-t^2\sin 4\theta\right]d\theta. \end{aligned} \tag{4.8}$$

Finally, it should be observed that one can use qip's to compose generating functions for different types of polynomials to construct examples involving mixed products. For example, if we select $f(x,t)$ as in A above and select

$$g(x,t)=e^{xt}\cosh(t\sqrt{x^2-1})=\sum_{n=0}^{\infty}\frac{T_n(x)}{n!}t^n$$

in which $T_n(x)$ denotes a Chebyshev polynomial of degree n (see [10]), then $f(x,\underline{t})\circ g(x,\underline{t})$ leads to the relation

$$\sum_{n=0}^{\infty}\frac{H_n(x)T_n(x)}{n!n!}t^{2n}=\frac{1}{2\pi}\cdot\int_0^{2\pi}\exp(3xt\cos\theta-t^2\cos 2\theta)\,h(x,t,\theta)\,d\theta$$

with

$$h(x,t,\theta)=\cos\alpha\cos\beta\cosh\gamma+\sin\alpha\sin\beta\sinh\gamma,$$

where

$$\alpha = xt\sin\theta - t^2\sin 2\theta,$$
$$\beta = t\sqrt{x^2-1}\,\sin\theta,$$
$$\gamma = t\sqrt{x^2-1}\,\cos\theta.$$

5. Some further relations. In the following, we first apply (1.1) and (1.4) to obtain transformations on the Lerch transcendent function. We then consider certain generalizations of integrals of the type involved in (1.1) and (1.4) that involve three functions. This will lead to some results on Bessel functions and some integral evaluations for Fourier coefficients.

A. *The Lerch transcendent function.* This function is defined by the series

$$\Phi(z,s,\alpha) = \sum_{n=0}^{\infty} (\alpha+n)^{-s} z^n, \qquad |z|<1. \tag{5.1}$$

For our purposes, we assume that $s>1$ and $\alpha \geqq 1$. Then the series in (5.1) converges when $|z|=1$. It follows that if $|x|\leqq 1$, $|y|\leqq 1$, then

$$\Phi(x,s_1,\alpha)\circ\Phi(y,s_2,\alpha) = \Phi(xy,s_1+s_2,\alpha) \qquad \text{for } s_1, s_2>1. \tag{5.2}$$

For the choice $\alpha=1$ and $s_1=s_2=s$, this becomes

$$\frac{1}{2\pi}\cdot\int_0^{2\pi} \Phi(e^{i\theta},s,1)\Phi(e^{-i\theta},s,1)\,d\theta = \zeta(2s), \tag{5.3}$$

where $\zeta(-)$ denotes the Riemann zeta function. For $\lambda\geqq 1$ and $(p,q)=1$, we also have

$$\Phi(x,s_1,\lambda q)_p{}^{\circ}{}_q\,\Phi(y,s_2,\lambda p) = \sum_{n=0}^{\infty}\left\{\frac{1}{(nq+\lambda q)^{s_1}}\frac{1}{(np+\lambda p)^{s_2}}\right\}x^{nq}y^{np}$$
$$= \frac{1}{q^{s_1}p^{s_2}}\sum_{n=0}^{\infty}\left\{\frac{1}{(n+\lambda)^{(s_1+s_2)}}\right\}(x^qy^p)^n$$

or

$$\Phi(x,s_1,\lambda q)_p{}^{\circ}{}_q\,\Phi(y,s_2,\lambda p) = \frac{1}{q^{s_1}p^{s_2}}\Phi(x^qy^p,s_1+s_2,\lambda). \tag{5.4}$$

B. *Some generalizations.* Let $f(z)$, $g(z)$, and $h(z)$ be analytic in D and consider the integral

$$I = \frac{1}{2\pi}\cdot\int_0^{2\pi}\left[f(xe^{pi\theta})g(ye^{\pm qi\theta})h(ze^{-ri\theta})\right]d\theta \tag{5.5}$$

where f and g have expansions as in the introduction and $h(z)=\sum_{n=0}^{\infty}c_nz^n$. Then, by replacing f, g, and h in I by their expansions, we obtain

$$I = \sum_{j,k,l=0}^{\infty} a_jb_kc_lx^jy^kz^l\frac{1}{2\pi}\cdot\int_0^{2\pi} e^{i(pj\pm qk-rl)\theta}\,d\theta. \tag{5.6}$$

The integrals in this sum vanish unless the summation indices satisfy the following Diophantine equation:

$$pj\pm qk-rl=0, \qquad j,k,l\geqq 0. \tag{5.7}$$

If one is able to solve for, say, the values of l in terms of the set of values for j and k, then the triple sum in (5.6) reduces to a double sum. If a partial summation can be carried out on this resulting double sum, then we can use (5.6) along with (5.5) to obtain further integral relations or evaluations for sums of special functions.

To avoid the complications of solving a general linear Diophantine equation, such as (5.7), in this paper we consider some specific choices for p,q,r. Suppose we first select $p=q=r=1$ and use the minus sign preceding the q in (5.7). Then we get $j=k+l$ so that (5.6) becomes

$$I=\sum_{k,l=0}^{\infty} a_{k+l}b_k c_l x^{k+l}y^k z^l. \tag{5.8}$$

As an interesting special case of (5.8), suppose we choose $f(z)=g(z)=h(z)=e^z$. Then the double sum I, with $x=y=z$, becomes

$$I=\sum_{k=0}^{\infty}\left(\sum_{l=0}^{\infty}\frac{1}{l!(l+k)!}z^{2l+k}\right)\frac{z^k}{k!}=\sum_{k=0}^{\infty}\frac{z^k}{k!}I_k(2z). \tag{5.9}$$

By reducing the integral in (5.5) to real form in θ using this choice of functions, we finally get

$$\sum_{k=0}^{\infty}\frac{z^k I_k(2z)}{k!}=\frac{1}{2\pi}\cdot\int_0^{2\pi} e^{3z\cos\theta}\cos(z\sin\theta)\,d\theta. \tag{5.10}$$

One can also use this method with $p=q=r=1$ and a plus sign in front of the q in (5.7) to deduce

$$\sum_{l=0}^{\infty}\frac{(-1)^l t^l}{l!}J_l(2x)=\frac{1}{2\pi}\cdot\int_0^{2\pi} e^{-t\cos\theta}\cos[(2x+t)\sin\theta]\,d\theta. \tag{5.11}$$

As a final example, suppose we select $f(x)=e^x$, $g(y)=h(y)=(1-y)^{-\lambda}$ and $p=q=r=1$ with a plus sign preceding the q in (5.7). The sum in (5.6) reduces to

$$\begin{aligned} I&=\sum_{j,k=0}^{\infty}\frac{1}{j!}\frac{(\lambda)_k}{k!}\frac{(\lambda)_{j+k}}{(k+j)!}x^j y^{2k+j} \qquad \text{(by (4.5b))}\\ &=\sum_{j=0}^{\infty}\frac{(xy)^j}{j!}\frac{(\lambda)_j}{(1)_j}\left(\sum_{k=0}^{\infty}\frac{(\lambda)_k(\lambda+j)_k}{k!(1+j)_k}y^{2k}\right) \qquad (5.12)\\ &=\frac{1}{\Gamma(\lambda)}\sum_{j=0}^{\infty}\frac{\Gamma(\lambda+j)(xy)^j}{(j!)^2}\,{}_2F_1(\lambda,\lambda+j;1+j;y^2), \end{aligned}$$

where $\Gamma(-)$ denotes the usual gamma function. Further, since $g(ye^{i\theta})h(ye^{-i\theta})=(1-2y\cos\theta+y^2)^{-\lambda}$, the corresponding integral (5.5) with integrand in real form in θ becomes

$$I=\frac{1}{2\pi}\cdot\int_0^{2\pi}\frac{e^{x\cos\theta}\cos(x\sin\theta)}{(1-2y\cos\theta+y^2)^{\lambda}}\,d\theta. \tag{5.13}$$

Now, select $\lambda = 1$. Then $\Gamma(\lambda + j) = j!$ and ${}_2F_1(\lambda, \lambda + j; 1 + j; y^2) = 1/(1 - y^2)$. The equality of (5.12) and (5.13) leads to the relation

$$\frac{1}{2\pi} \cdot \int_0^{2\pi} \frac{e^{x\cos\theta} \cos(x \sin\theta)(1 - y^2)}{(1 - 2y\cos\theta + y^2)} \, d\theta = e^{xy}. \tag{5.14}$$

The expansion of both sides of this in powers of y and a comparison of their corresponding coefficients yields the following integral evaluations:

$$\begin{aligned} &\frac{1}{2\pi} \cdot \int_0^{2\pi} e^{x\cos\theta} \cos(x \sin\theta) \, d\theta = 1, \\ &\frac{1}{2\pi} \cdot \int_0^{2\pi} e^{x\cos\theta} \cos(x \sin\theta) \cos n\theta \, d\theta = \frac{x^n}{n!}, \qquad n \geqq 1. \end{aligned} \tag{5.15}$$

REFERENCES

[1] L. R. BRAGG, *Hypergeometric operator series and related partial differential equations*, Trans. AMS, 143 (1969), pp. 319–336.

[2] ———, *Products of certain Hermite polynomials: associated relations*, Bull. UMI, 3 (1968), pp. 347–355.

[3] L. R. BRAGG AND J. W. DETTMAN, *An operator calculus for related partial differential equations*, J. Math. Anal. Appl., 22 (1968), pp. 459–467.

[4] R. V. CHURCHILL, *Fourier Series and Boundary Value Problems*, McGraw-Hill, New York, 1941.

[5] P. DUREN, *Theory of H^p Spaces*, Academic Press, New York, 1960.

[6] P. L. DURAN, B. W. ROMBERG, AND A. L. SHIELDS, *Linear functionals on H^p-spaces with $0 < p < 1$*, J. Reine Angew. Math., 238 (1969), pp. 32–60.

[7] H. W. GOULD AND A. T. HOPPER, *Operational formulas connected with two generalizations of Hermite polynomials*, Duke Math. J., 29 (1962), pp. 51–64.

[8] M. L. J. HAUTUS AND D. A. KLARNER, *The diagonal of a double power series*, Duke Math. J., 38 (1971), pp. 229–235.

[9] K. HOFFMAN, *Banach Spaces of Analytic Functions*, Prentice-Hall, Englewood Cliffs, NJ, 1962.

[10] W. MAGNUS, F. OBERHETTINGER AND R. SONI, *Formulas and Theorems for the Special Functions of Mathematical Physics*, Springer-Verlag, New York, 1966.

[11] E. RAINVILLE, *Special Functions*, Macmillan, New York, 1960.

[12] W. RUDIN, *Real and Complex Analysis*, McGraw-Hill, New York, 1966.

[13] E. C. TITCHMARSH, *The Theory of Functions*, Oxford Univ. Press, Oxford, 1949.

SIAM J. MATH. ANAL.
Vol. 17, No. 1, January 1986

020

UNIVALENCE CONSTRAINTS ON THE SCHWARZ–CHRISTOFFEL PARAMETERS*

JOHN A. PFALTZGRAFF†

Dedicated to Professor A. W. Goodman on the occasion of his seventieth birthday.

Abstract. The Area Theorem and coefficient inequalities for univalent functions are used to derive inequality constraints on the accessory parameters in the Schwarz–Christoffel formula for the conformal mapping of the unit disk onto the interior or onto the exterior of a polygon. The constraints restrict the choice of prevertices and are necessary for the univalence of the mapping. The conditions are explicit, easy to compute and are applicable to the numerical computation of the Schwarz–Christoffel map. The relative effectiveness of the constraints is illustrated by elementary examples.

Key words. accessory parameters, area theorem, conformal map, exterior mapping, prevertices, univalent, Schwarz–Christoffel formula

1. Introduction. In general it is very difficult to determine $z_k = e^{i\theta_k}$ $(k=1,\cdots,N)$, the *prevertices* (accessory parameters), in the Schwarz–Christoffel formula,

$$f(z) = C\int_0^z \prod_{k=1}^{N} (\zeta - z_k)^{-\beta_k} d\zeta + f(0), \tag{1.1}$$

$$\theta_1 < \theta_2 < \cdots < \theta_N < \theta_1 + 2\pi, \quad -1 \leqq \beta_k \leqq 3, \quad \sum_{k=1}^{N} \beta_k = 2, \tag{1.2}$$

for the conformal map of the unit disk $E = \{z: |z|<1\}$ onto the interior of a polygon P with interior angles $\pi(1-\beta_k)$ at the vertices $w_k = f(z_k)$, $k=1,\cdots,N$. When $-1 \leqq \beta_k < 1$ the vertex w_k is finite and $\pi\beta_k$ measures the turning of the tangent at w_k as the boundary of P is traversed in a counterclockwise direction. Vertices at ∞ correspond to the β_k with $1 \leqq \beta_k \leqq 3$ and the usual interpretation of interior and exterior angle at such a vertex [2, Thm. 5.12 e], [6, p. 83].

It is well known that formula (1.1) forces the image of the unit circle, $f(|z|=1)$, to be a polygonal path with vertices $w_k = f(z_k)$ and the specified exterior angles $\pi\beta_k$ at each vertex. The main difficulty in applying (1.1) is the choosing of C and the prevertices so that f will be univalent in E and P will have the correct side lengths (see [6, pp. 83–84]).

In this note we use results from univalent function theory to derive inequality constraints on the prevertices that are necessary for f in (1.1) to be univalent in E. We also give analogous conditions for the Schwarz–Christoffel mapping of E onto the exterior of a polygon. The accessory parameter problem is the most difficult and time consuming task in the numerical computation of the Schwarz–Christoffel transformation [6]. Although our results are extremely elementary, they provide explicit, computable constraints on the choice of accessory parameters.

*Received by the editors February 7, 1984.
† Department of Mathematics, University of North Carolina, Chapel Hill, North Carolina 27514.

2. The interior mapping. Since $z_k\bar{z}_k=1$ in (1.1) we may rewrite that formula as

$$f(z)=C\int_0^z \prod_{k=1}^{N}(1-\zeta\bar{z}_k)^{-\beta_k}\,d\zeta+f(0) \tag{2.1}$$

with a new constant C. Our necessary conditions for univalence of f involve the weighted power sums

$$s_n=\sum_{k=1}^{N}\beta_k\bar{z}_k^{n+1}, \qquad n=0,1,\cdots, \tag{2.2}$$

that appear as the Taylor coefficients in

$$\frac{f''(z)}{f'(z)}=\sum_{k=1}^{N}\frac{\beta_k\bar{z}_k}{1-z\bar{z}_k}=\sum_{n=0}^{\infty}s_nz^n, \qquad z\in E. \tag{2.3}$$

Note that (2.3) is independent of $f(0)$ and the scale factor $C=f'(0)$. Indeed, if $f(z)$ is univalent in E then

$$F(z)=\frac{f(z)-f(0)}{C}=z+\sum_{n=2}^{\infty}a_nz^n, \qquad z\in E, \tag{2.4}$$

belongs to the class S of normalized univalent functions and

$$\frac{F''(z)}{F'(z)}=\frac{f''(z)}{f'(z)}=\sum_{n=0}^{\infty}s_nz^n, \qquad z\in E. \tag{2.5}$$

THEOREM 1. *Let $z_k=e^{i\theta_k}$ and $\beta_k(k=1,\cdots,N)$ be given subject to conditions* (1.2). *If the function $f(z)$ defined by* (2.1) *is univalent in E then*

$$|s_0|=\left|\sum_{k=1}^{N}\beta_kz_k\right|\leqq 4, \tag{2.6}$$

$$|s_1+s_0^2|=\left|\sum_{k=1}^{N}\beta_kz_k^2+\left(\sum_{k=1}^{N}\beta_kz_k\right)^2\right|\leqq 18 \tag{2.7}$$

and

$$\left|s_1-\frac{1}{2}s_0^2\right|=\left|\sum_{k=1}^{N}\beta_kz_k^2-\frac{1}{2}\left(\sum_{k=1}^{N}\beta_kz_k\right)^2\right|\leqq 6. \tag{2.8}$$

Proof. Using the series expansion (2.4) in (2.5) we have

$$\sum_{n=0}^{\infty}(n+1)(n+2)a_{n+2}z^n=\left(1+\sum_{n=1}^{\infty}(n+1)a_{n+1}z^n\right)\sum_{n=0}^{\infty}s_nz^n$$

and the system

$$(n+1)(n+2)a_{n+2}=s_n+\sum_{k=1}^{n}(k+1)a_{k+1}s_{n-k}, \qquad n=0,1,\cdots, \tag{2.9}$$

which can be solved recursively to express a_n in terms of the s_k $(k\leqq n-2)$. The first two formulas are

$$2a_2=s_0, \qquad 6a_3=s_1+s_0^2. \tag{2.10}$$

If f is univalent in E then F belongs to S and therefore

$$|a_2| \leqq 2, \quad |a_3| \leqq 3, \quad |a_3 - a_2^2| \leqq 1, \tag{2.11}$$

[5, p. 213], [1, p. 94], which with (2.10) give (2.6), (2.7), and (2.8) respectively.

Clearly one can continue the foregoing method of proof with (2.9) and the sharp coefficient bounds $|a_n| \leqq n$ [7] to extend (2.6)–(2.8) to an infinite sequence of conditions on the s_n necessary for f to be univalent. Our second example below shows that (2.8) is more sensitive to nonunivalence of f than either (2.6) or (2.7). Note that $|a_3 - a_2^2| \leqq 1$ is simply the inequality $|b_1| \leqq 1$, obtained by applying the area theorem, $\sum_{n=1}^{\infty} n|b_n|^2 \leqq 1$ [5, p. 210] to

$$\frac{1}{F(z)} = \frac{1}{z} + b_0 + b_1 z + \cdots. \tag{2.12}$$

Thus it may be more fruitful to use the series expansion (2.4) in (2.12) to obtain the system

$$-b_n = a_{n+2} + \sum_{k=1}^{n-1} b_k a_{n+1-k}, \qquad n = 0, 1, \cdots, \tag{2.13}$$

solve (2.9) and (2.13) recursively to express b_n $(n = 1, \cdots, m)$ in terms of the s_k, and then apply the area theorem inequalities $\sum_{n=1}^{m} n|b_n|^2 \leqq 1$, $m = 1, 2, \cdots$.

Example 1. Let $f(z)$ be defined by (2.1) with $N = 4$, $(\beta_1, \beta_2, \beta_3, \beta_4) = (-\frac{3}{4}, -\frac{3}{4}, \frac{3}{4}, \frac{11}{4})$, $z_1 = (-5 + i\sqrt{11})/6, z_2 = \bar{z}_1, z_3 = -i, z_4 = 1$. Then f is not univalent since

$$\left| \sum_{k=1}^{4} \beta_k z_k \right| = \left| 4 - i\frac{3}{4} \right| > 4,$$

violating condition (2.6).

Example 2. Let $f(z)$ be defined by (2.1) with $N = 4$, β_k $(k = 1, 2, 3, 4)$ the same as in Example 1, $z_1 = (-47 + i\sqrt{1391})/60$, $z_2 = \bar{z}_1$, $z_3 = -i$, $z_4 = 1$. Then easy calculations show that both (2.6) and (2.7) are satisfied (3.9960... for (2.6) and 17.521... for (2.7)), but (2.8) is violated (6.470... is the value of (2.8)).

Remarks. Condition (2.6) is the unit disk version of E. Johnston's condition (3) (with $z_0 = i$) in [3] for the Schwarz–Christoffel mapping of the upper half-plane onto a polygon. We are able to construct nonunivalent examples with only four vertices (rather than six as in [3]) because we permit vertices at ∞ with exterior angles as large as 3π. Our formulation for the unit disk eliminates the need to translate results about the class S from E to the upper half-plane [3, Lemma, p. 702]. Furthermore the weighted power sums s_n are easy to compute, and they have a nice geometrical interpretation as moments of a distribution of weights β_k $(-1 \leqq \beta_k \leqq 3)$, on the unit circle.

3. The exterior mapping. If $g(z)$ is analytic in $0 < |z| < 1$ with a simple pole at 0, and if g maps E conformally onto the exterior of a bounded polygon P then g is given by the Schwarz–Christoffel formula

$$g(z) = C \int_a^z \prod_{k=1}^{N} (1 - \zeta \bar{z}_k)^{-\beta_k} \zeta^{-2} \, d\zeta + C_1, \tag{3.1}$$

$z_k = e^{i\theta_k}$, $\theta_1 < \theta_2 < \cdots < \theta_N < \theta_1 + 2\pi$ and

$$-1 \leqq \beta_k < 1, \quad \sum_{k=1}^{N} \beta_k = -2, \quad \sum_{k=1}^{N} \beta_k \bar{z}_k = 0, \tag{3.2}$$

[4, p. 331], [2, pp. 413–414]. The points $w_k = g(z_k)$, $k = 1, \cdots, N$, are the vertices of P, and since g reverses the boundary orientation, $\pi\beta_k$ measures the turning of the tangent at w_k as the boundary of P is traversed in a *clockwise* direction. The condition $\sum_{k=1}^{N} \beta_k \bar{z}_k = 0$ insures that $g'(z)$ has zero residue at the origin and hence that $g(z)$ is single valued. Indeed, if this condition holds then $g(z)$ has the Laurent expansion

$$g(z) = C\left(-\frac{1}{z} + \sum_{n=0}^{\infty} b_n z^n\right), \qquad 0 < |z| < 1, \tag{3.3}$$

and

$$z^2 g'(z) = C \prod_{k=1}^{N} (1 - z\bar{z}_k)^{-\beta_k} = C\left(1 + \sum_{n=1}^{\infty} n b_n z^{n+1}\right),$$

with

$$\operatorname{Res}(g'(z); 0) = C \sum_{k=1}^{N} \beta_k \bar{z}_k = 0.$$

The weighted power sums s_n defined by (2.2) now appear as the Taylor coefficients in

$$\begin{aligned} \frac{(z^2 g'(z))'}{z^2 g'(z)} &= \sum_{k=1}^{N} \frac{\beta_k \bar{z}_k}{1 - z\bar{z}_k} = \sum_{n=1}^{\infty} s_n z^n \\ &= \left(\sum_{n=1}^{\infty} n(n+1) b_n z^n\right) \Big/ \left(1 + \sum_{n=1}^{\infty} n b_n z^{n+1}\right), \qquad z \in E. \end{aligned} \tag{3.4}$$

THEOREM 2. *Let $z_k = e^{i\theta_k}$ and β_k $(k = 1, \cdots, N)$ be given satisfying conditions* (3.2). *If the function $g(z)$ defined by* (3.1) *is univalent in E then*

$$|s_1| = \left| \sum_{k=1}^{N} \beta_k z_k^2 \right| \leqq 2. \tag{3.5}$$

Proof. Using the series expansions in (3.4) we have $2b_1 = s_1$ and

$$n(n+1) b_n = s_n + \sum_{k=1}^{n-2} k b_k s_{n-1-k}, \qquad n = 2, 3, \cdots. \tag{3.6}$$

If g is univalent in E then by the area theorem $\sum_{n=1}^{\infty} n|b_n|^2 \leqq 1$. In particular $|b_1| \leqq 1$ which yields (3.5).

Again it is clear that additional univalence constraints on the s_n can be obtained by solving (3.6) recursively for b_n in terms of the s_k and then applying the area theorem. The process is simpler for the exterior mapping than for the interior mapping where there are two systems, (2.9) and (2.13), to solve.

Example 3. Let $g(z)$ be defined by (3.1) with $N=5$, $(\beta_1,\beta_2,\beta_3,\beta_4,\beta_5)=(-\frac{9}{10},-\frac{1}{2},-\frac{1}{2},-\frac{9}{10},\frac{4}{5})$, $z_1=\bar{z}_4=(4+i\sqrt{65})/9$, $z_2=\bar{z}_3=i$, $z_5=1$. Then the β_k and z_k satisfy conditions (3.2), but g is not univalent since

$$b_1=\frac{1}{2}\sum_{k=1}^{5}\beta_k z_k^2=1+\frac{4}{9}>1.$$

REFERENCES

[1] P. L. DUREN, *Univalent Functions*, Springer-Verlag, New York, 1983.

[2] P. HENRICI, *Applied and Computational Complex Analysis*, Vol. 1, John Wiley, New York, 1974.

[3] E. JOHNSTON, *A "counterexample" for the Schwarz–Christoffel transform*, Amer. Math. Monthly, 90 (1983), pp. 701–703.

[4] A. I. MARKUSHEVICH, *Theory of Functions of a Complex Variable*, Vol. 3, Prentice-Hall, Englewood Cliffs, NJ, 1967.

[5] Z. NEHARI, *Conformal Mapping*, McGraw-Hill, New York, 1952.

[6] L. N. TREFETHEN, *Numerical computation of the Schwarz–Christoffel transformation*, SIAM J. Sci. Stat. Comp., 1 (1980), pp. 82–102.

[7] L. DE BRANGES, *A proof of the Bieberbach conjecture*, Acta Math., 154 (1985), pp. 137–152.

SIAM J. MATH. ANAL.
Vol. 17, No. 1, January 1986

021

CRITICAL VALUES AND REPRESENTATION OF FUNCTIONS BY MEANS OF COMPOSITIONS*

Y. YOMDIN†

Abstract. The structure of the set of critical values of a composition of differentiable mappings is studied. On this base some explicit examples of differentiable functions, not representable by compositions of certain types, are given.

The question of a representability of functions of a given class by means of compositions of functions, belonging to some other given classes, has been studied in many publications (see e.g. [1], [2], [3], [4]). The known results on a nonrepresentability are mostly based on considerations of a "massiveness" of corresponding subsets in a suitable functional space. Therefore, showing the existence of nonrepresentable functions, they do not give explicit examples.

In this paper we consider some special question of the above type, which roughly can be formulated as follows: what functions of a given smoothness and a given number of variables can be represented in a form $f=g\circ h$, where g depends on a smaller number of variables than f (and, possibly, has a lower smoothness), while the smoothness of h is higher than the smoothness of f.

We find an upper bound for the entropy dimension of the set of critical values of the function, represented in such a form. Using this bound we give the necessary condition for the representability, and some examples of nonrepresentable functions.

The result of Theorem 1 below can be considered as a generalization to the case of composed mappings of the sharp estimates for the entropy dimension of critical values of differentiable mappings, obtained in [6].

Let f: $U\to R^m$ be a continuously differentiable mapping of an open domain $U\subset R^n$. For $\gamma\geqq 0$ define the set of "γ-near-critical" points of f, $\Sigma(f,\gamma)$, by $\Sigma(f,\gamma)=\{x\in U/\|df(x)\|\leqq\gamma\}$, and let $\Delta(f,\gamma)=f(\Sigma(f,\gamma))$ be the corresponding set of "γ-near-critical" values of f. The set of critical points $\Sigma(f,0)$ we denote by $\Sigma(f)$ and the set of critical values $\Delta(f,0)$ by $\Delta(f)$.

For a compact domain $D\subset R^n$ let $C^k(D,m)$ denote the space of mappings f: $D\to R^m$, which can be extended to a k times continuously differentiable mapping of some open neighborhood of D. For $f\in C^k(D,m)$ let

$$M_i(f)=\max_{y\in D}\|d^i f(y)\|,\qquad i=0,\cdots,k.$$

(The Euclidean spaces R^q and the spaces of their linear and multilinear mappings are considered with the usual Euclidean norms.)

We recall the definition of the entropy dimension: for a bounded subset $A\subset R^m$ and $\varepsilon>0$ let $M(\varepsilon,A)$ be the minimal number of balls of radius ε, covering A. Then the entropy dimension $\dim_e A$ is defined by

$$\dim_e A=\inf\left\{\beta\geqq 0/\exists K,\forall\varepsilon,1\geqq\varepsilon>0,M(\varepsilon,A)\leqq K\left(\frac{1}{\varepsilon}\right)^{\beta}\right\}.$$

(See e.g. [3], [6] for some properties of $M(\varepsilon,A)$ and of the entropy dimension.)

* Received by the editors November 29, 1983, and in revised form August 13, 1984.
† Department of Mathematics, Ben-Gurion University of the Negev, Beer-Sheva 84120, Israel.

Now let $B^\alpha \subset R^{n_\alpha}$ be closed balls of radii ρ_α, respectively, $\alpha = 1, \cdots, s+1$, and let f: $B^1 \to B^{s+1}$ be given as a composition $f = f^s \circ f^{s-1} \circ \cdots \circ f^1$, where f^α: $B^\alpha \to B^{\alpha+1}$, $f^\alpha \in C^{k_\alpha}(B^\alpha, n_{\alpha+1})$, $k_\alpha \geqq 2$, $\alpha = 1, \cdots, s$.

Below we assume that the dimensions n_α do not increase: $n_1 \geqq n_2 \geqq \cdots \geqq n_s \geqq n_{s+1}$. (The problem of a factorization through lower-dimensional spaces has an essentially different nature.) Without loss of generality we can also assume that $k_1 > k_2 > \cdots > k_s$. Indeed, one can easily prove that if $k_2 \geqq k_1$ and $n_1 \geqq n_2 \geqq n_3$, then all the mappings f: $B^1 \to B^3$, belonging to $C^{k_1}(B^1, n_3)$, and only these mappings, are representable as $f = f^2 \circ f^1$, f^α: $B^\alpha \to B^{\alpha+1}$, $f^\alpha \in C^{k_\alpha}(B^\alpha, n_{\alpha+1})$, $\alpha = 1, 2$.

Let f be represented as a superposition of s mappings as above. We call $S = (n_1, k_1, n_2, k_2, \cdots, n_s, k_s)$ the diagram of the representation and say that f is representable with the diagram S.

For a given diagram S define $\sigma(S)$ as

$$\frac{n_1 - n_2}{k_1 - 1} + \frac{n_2 - n_3}{k_2 - 1} + \cdots + \frac{n_{s-1} - n_s}{k_{s-1} - 1} + \frac{n_s}{k_s - 1}.$$

THEOREM 1. *If f is representable with the diagram S, then*

$$\dim_e \Delta(f) \leqq \sigma(S).$$

Proof. Fix some $\varepsilon > 0$, $\varepsilon \leqq 1$. Below we denote by K_j some constants, depending only on the set of data $Q = \{\rho_\alpha, n_\alpha, k_\alpha, M_i(f^\alpha), \alpha = 1, \cdots, s+1, i = 0, \cdots, k_\alpha\}$, but not on ε.

For each $\alpha = 1, \cdots, s$ let $r_\alpha = \varepsilon^{1/(k_\alpha - 1)}$. Consider in each $B^\alpha \subset R^{n_\alpha}$ the points with the coordinates of the form $mr_\alpha/\sqrt{n_\alpha}$, $m \in Z$, and denote these poins by x_β^α, $1 \leqq \beta \leqq d_\alpha$. The balls B_β^α of the radius r_α, centered at x_β^α, cover B^α.

Let P_β^α be the Taylor polynomial of order k_α of the mapping f^α at the point x_β^α.

For any s-tuple $(\beta_1, \cdots, \beta_s)$, $1 \leqq \beta_\alpha \leqq d_\alpha$, denote by $P_{\beta_1, \cdots, \beta_s}$ the polynomial mapping

$$P_{\beta_s}^s \circ P_{s-1}^{s-1} \circ \cdots \circ P_{\beta_1}^1 \colon R^{n_1} \to R^{n_{s+1}}.$$

We also denote by $D_{\beta_1, \cdots, \beta_s}$ the set

$$\left\{ x \in B_{\beta_1}^1 / f^\alpha \circ f^{\alpha-1} \circ \cdots \circ f^1(x) \in B_{\beta_{\alpha+1}}^{\alpha+1}, \alpha = 1, \cdots, s-1 \right\}.$$

LEMMA 2. *For any s-tuple $(\beta_1, \cdots, \beta_s)$ and for any $x \in D_{\beta_1, \cdots, \beta_s}$,*
(i) $\|f(x) - P_{\beta_1, \cdots, \beta_s}(x)\| \leqq K_1 \varepsilon$,
(ii) $\|df(x) - dP_{\beta_1, \cdots, \beta_s}(x)\| \leqq K_1 \varepsilon$.

Proof. By induction on α. Denote $f^\alpha \circ f^{\alpha-1} \circ \cdots \circ f^1$ by F^α, and $P_{\beta_\alpha}^\alpha \circ \cdots \circ P_{\beta_1}^1$ by Q^α and assume that (i) and (ii) are satisfied for $F^{\alpha-1}$, $Q^{\alpha-1}$ with the constant K_1^α.

By the choice of r_α and by the Taylor formula we have for any $y \in B_{\beta_\alpha}^\alpha$:

$$\left\| f^\alpha(y) - P_{\beta_\alpha}^\alpha(y) \right\| \leqq K_2 r_\alpha^k = K_2 \varepsilon^{k_\alpha/(k_\alpha - 1)} \leqq K_2 \varepsilon,$$

$$\left\| df^\alpha(y) - dP_{\beta_\alpha}^\alpha \right\| \leqq K_3 r_\alpha^{k_\alpha - 1} = K_3 \varepsilon.$$

Now denote $F^{\alpha-1}(x)$ by y_1, $Q^{\alpha-1}(x)$ by y_2. Since $x \in D_{\beta_1, \cdots, \beta_s}$, $y_1 \in B_{\beta_\alpha}^\alpha$. We have:

$$\begin{aligned} \|F^\alpha(x) - Q^\alpha(x)\| &= \left\| f^\alpha(y_1) - P_{\beta_\alpha}^\alpha(y_2) \right\| \\ &\leqq \left\| f^\alpha(y_1) - P_{\beta_\alpha}^\alpha(y_1) \right\| + \left\| P_{\beta_\alpha}^\alpha(y_1) - P_{\beta_\alpha}^\alpha(y_2) \right\| \\ &\leqq K_2 \varepsilon + M_1\left(P_{\beta_\alpha}^\alpha\right) \|y_1 - y_2\| \leqq K_2 \varepsilon + K_4 K_1^{\alpha - 1} \varepsilon = K_5 \varepsilon. \end{aligned}$$

In a similar way, for the first derivatives we have:

$$\begin{aligned}\|dF^{\alpha}(x)-dQ^{\alpha}(x)\| &= \|df^{\alpha}(y_1)\circ dF^{\alpha-1}(x)-dP^{\alpha}_{\beta_\alpha}(y_2)\circ dQ^{\alpha-1}(x)\| \\ &\leqq \|df^{\alpha}(y_1)-dP^{\alpha}_{\beta_\alpha}(y_1)\|\|dF^{\alpha-1}(x)\| \\ &\quad + \|dP^{\alpha}_{\beta_\alpha}(y_1)\|\|dF^{\alpha-1}(x)-dQ^{\alpha-1}(x)\| \\ &\quad + \|dP^{\alpha}_{\beta_\alpha}(y_1)-dP^{\alpha}_{\beta_\alpha}(y_2)\|\|dQ^{\alpha-1}(x)\| \\ &\leqq K_3\varepsilon K_6 + K_7K_1^{\alpha-1}\varepsilon + M_2\left(P^{\alpha}_{\beta_\alpha}\right)K_1^{\alpha-1}\varepsilon K_8 = K_9\varepsilon.\end{aligned}$$

Denoting $\max(K_5, K_9)$ by K_1^{α}, we obtain the required inequalities, with $K_1 = K_1^s$.

For any s-tuple $(\beta_1,\cdots,\beta_s)$ let $\Sigma_{\beta_1,\cdots,\beta_s} = \Sigma(f)\cap D_{\beta_1,\cdots,\beta_s} \subset B^1$ and let $\Delta_{\beta_1,\cdots,\beta_s} = f(\Sigma_{\beta_1,\cdots,\beta_s})$.

LEMMA 3. *For any $(\beta_1,\cdots,\beta_s)$, $\Delta_{\beta_1,\cdots,\beta_s}$ can be covered by K_{10} balls of radius ε.*

Proof. By Lemma 2 ii, $\Sigma_{\beta_1,\cdots,\beta_s} \subset \tilde{\Sigma}(P_{\beta_1,\cdots,\beta_s}, K_1\varepsilon)$. By i, $\Delta_{\beta_1,\cdots,\beta_s}$ is thus contained in a $K_1\varepsilon$-neighborhood of $\Delta(P_{\beta_1,\cdots,\beta_s}, K_1\varepsilon)$. Now by [6, Cor. 2.14], this last set, being the set of near critical values of the polynomial mapping $P_{\beta_1,\cdots,\beta_s}$ on the ball B^1 of radius ρ_1, can be covered by N balls of radius $\rho_1 K_1\varepsilon$, where $N = N(n_1, n_{s+1}, k_1\cdot k_2\cdots k_s)$ depends only on the dimensions n_α and on the degrees of differentiability k_α. Hence $\Delta_{\beta_1,\cdots,\beta_s}$ can be covered by the same number of balls of radius $(\rho_1+1)K_1\varepsilon$, or by $N[\sqrt{n_{s+1}}(2\rho_1+2)K_1]^{n_{s+1}} = K_{10}$ balls of radius ε.

Of course, all the sets $\Delta_{\beta_1,\cdots,\beta_s}$, $1 \leqq \beta_\alpha \leqq d_\alpha$, cover $\Delta(f)$. But in fact many of these sets are empty. So to prove Theorem 1 it remains to estimate the number of nonempty $\Delta_{\beta_1,\cdots,\beta_s}$, which, in turn, does not exceed the number of nonempty $D_{\beta_1,\cdots,\beta_s}$.

LEMMA 4. *Let $\beta_1,\cdots,\beta_{\alpha-1}$ be fixed, $\alpha = 1,\cdots,s$. Then the number of the indices β_α, for which $D_{\beta_1,\cdots,\beta_{\alpha-1},\beta_\alpha,\beta_{\alpha+1},\cdots,\beta_s}$ is not empty for some $\beta_{\alpha+1},\cdots,\beta_s$, does not exceed $K_{11}(1/\varepsilon)^{-n_\alpha/(k_{\alpha-1}-1)+n_\alpha/(k_\alpha-1)}$ (where $k_0 \overset{\text{def}}{=} \infty$).*

Proof. By definition, $D_{\beta_1,\cdots,\beta_s} = \{x \in B^1_{\beta_1}/F^{\alpha-1}(x) \in B^{\alpha}_{\beta_\alpha}, \alpha = 1,\cdots,s\}$. Hence $F^{\alpha-1}(D_{\beta_1,\cdots,\beta_s}) \subset B^{\alpha}_{\beta_\alpha} \cap f^{\alpha-1}(B^{\alpha-1}_{\beta_{\alpha-1}})$. Now $B^{\alpha-1}_{\beta_{\alpha-1}}$ is a ball of radius $r_{\alpha-1}$, and therefore $f^{\alpha-1}(B^{\alpha-1}_{\beta_{\alpha-1}})$ is contained in some ball B in R^{n_α} of radius $M_1(f^{\alpha-1})r_{\alpha-1} \leqq K_{12}r_{\alpha-1}$. But clearly B has nonempty intersection with not more than

$$\left(2\sqrt{n_\alpha}K_{12}r_{\alpha-1}/r_\alpha\right)^{n_\alpha} = K_{11}\left(\frac{1}{\varepsilon}\right)^{n_\alpha/(k_\alpha-1)-n_\alpha/(k_{\alpha-1}-1)}$$

balls B^{α}_{β}.

Thus the number of nonempty $D_{\beta_1,\cdots,\beta_s}$ is bounded by

$$K_{11}^s\left(\frac{1}{\varepsilon}\right)^{n_1/(k_1-1)-n_2/(k_1-1)+n_2/(k_2-1)-\cdots-n_s/(k_{s-1}-1)+n_s/(k_s-1)} = K_{13}\left(\frac{1}{\varepsilon}\right)^{\sigma(S)}.$$

Since by Lemma 3 each $\Delta_{\beta_1,\cdots,\beta_s}$ can be covered by K_{10} balls of radius ε, we have $M(\varepsilon, \Delta(f)) \leqq K_{10}K_{13}(1/\varepsilon)^{\sigma(S)}$. Theorem 1 is proved.

Theorem 1 can be applied to the representability question in the following way: clearly, any mapping f, representable with the diagram $S = (n_1, k_1, \cdots, n_s, k_s)$, $n_1 \geqq n_2 \geqq \cdots \geqq n_s$, $k_1 > k_2 > \cdots > k_s$, is at least k_s-smooth. On the other hand, any k_1-smooth mapping is representable with the diagram S by the remark above. Hence, the question is nontrivial for mappings f: $B^1 \to B^{s+1}$ of smoothness q, $k_1 > q > k_s$.

Now we use the sharp estimate of the entropy dimension of the set of critical values, obtained in [6]. By [6, Thm. 5.6], for any $f \in C^q(B^1, n_{s+1})$, $\dim_e \Delta(f) \leqq n_1/q$, and for any $\eta < n_1/q$ there are mappings $f \in C^q(B^1, n_{s+1})$ with $\dim_e \Delta(f) > \eta$. Thus we have the following:

COROLLARY 5. *Let S be a given diagram and let $k_1 > q > k_s$. If $n_1/q > \sigma(S)$, then there are mappings $f \in C^q(B^1, n_{s+1})$, not representable with the diagram S.*

As a criterion of existence of nonrepresentable mappings, Corollary 5 is weaker than the result of Vitushkin, Kolmogorov and Tikhomirov [3], [4]. Indeed, by [3, Thm. XXVII], there are mappings in $C^q(B_1, n_{s+1})$, nonrepresentable with the diagram S, if $n_1/q > \max_\alpha(n_\alpha/k_\alpha)$. But easy computations show that always $\delta(S) \geqq \max_\alpha(n_\alpha/k_\alpha)$, and usually the strict inequality holds.

The reason is that the entropy dimension of critical values of a composition can be greater than that of each of the composed mappings, while the functional dimension of the class of representable mappings (considered in [3], [4]), is equal to the maximal functional dimension of classes, participating in the representation.

But this functional dimension, as an invariant of the whole class, does not allow to find explicit examples of nonrepresentable functions. On the contrary, the entropy dimension of critical values is the invariant of individual mappings, and the mappings with the "big" set of critical values can be built explicitly. Hence in any situation, covered by Corollary 5, we can find explicit examples of nonrepresentable mappings. In particular, let $B \subset R^n$ be the closed unit ball and let $h_n \in C^{n-1}(B, 1)$ be the Whitney function (see [5]) with $\Delta(f) = [0, 1]$. Since $\dim_e [0, 1] = 1$, we obtain

COROLLARY 6. *h_n cannot be represented with the diagram $S = (n, k_1, n_2, k_2)$, if*

$$\frac{n - n_2}{k_1 - 1} + \frac{n_2}{k_2 - 1} < 1.$$

In particular, $h_{10}(x_1, \cdots, x_{10})$ cannot be written as $h_{10}(x_1, \cdots, x_{10}) = \psi(y_1, \cdots, y_5)$, $y_i = \psi_i(x_1, \cdots, x_{10})$, $i = 1, \cdots, 5$, with $\psi \in C^7$ and $\psi_i \in C^{32}$.

There is some similarity in the properties of functions representable by means of compositions of smooth functions and maximum functions of smooth families (compare [7]), although the last class contains nondifferentiable functions.

It is interesting whether direct connections between these two classes can be found.

REFERENCES

[1] V. I. ARNOLD, *On functions of three variables*, Dokl. Akad. Nauk SSSR, 114 (1957), 679–681. (In Russian.)

[2] A. N. KOLMOGOROV, *On representation of continuous functions of several variables by superpositions of continuous functions of fewer variables*, Dokl. Akad. Nauk SSSR, 108 (1956), pp. 179–182; Amer. Math. Soc. Transl. (2) 17 (1961), pp. 369–373.

[3] A. N. KOLMOGOROV AND V. M. TIKHOMIROV, *ε-entropy and ε-capacity of sets in function spaces*, Uspekhi Mat. Nauk, 14, 2 (1959), pp. 3–86; Amer. Math. Soc. Transl. (2) 17 (1961), pp. 277–364.

[4] A. G. VITUSHKIN, *On representation of functions by means of superpositions and related topics*, Enseignement Math., 23 (1977), pp. 255–320.

[5] H. WHITNEY, *A function not constant on a connected set of critical points*, Duke Math. J., 1 (1935), pp. 514–517.

[6] Y. YOMDIN, *The geometry of critical and near-critical values of differentiable mappings*, Math. Ann., 264 (1983), pp. 495–515.

[7] ———, *On functions representable as a supremum of a family of smooth functions*, II, this Journal, to appear.

SIAM J. MATH. ANAL.
Vol. 17, No. 1, January 1986

022

POLYNOMIAL ANALOGUES OF PROLATE SPHEROIDAL WAVE FUNCTIONS AND UNCERTAINTY*

MARCI PERLSTADT†

Abstract. Slepian, Landau, and Pollak used prolate spheroidal wave functions to demonstrate how nearly "time" and "bandlimited" a square-integrable function can be. In this note we show how their results extend easily to cover orthogonal polynomial expansions. In particular, we study how close a square-integrable function can come to being a polynomial of degree $\leqq L$ and simultaneously to vanishing off some set $\mathscr{A}$.

1. Introduction. Let $f \in L^2(\mathbb{R})$ with Fourier transform $\hat{f}$. We say that f is time-limited to the set $\mathscr{A} \subset \mathbb{R}$ if $\chi_{\mathscr{A}} \cdot f = f$ and that f is bandlimited to the set $\mathscr{B} \subset \mathbb{R}$ if $\chi_{\mathscr{B}} \cdot \hat{f} = \hat{f}$ (here $\chi_{\mathscr{A}}$, $\chi_{\mathscr{B}}$ are the respective characteristic functions of the sets $\mathscr{A}$, $\mathscr{B}$). It is well known that f cannot be simultaneously time and bandlimited and the extend to which f can be "approximately" time and bandlimited has been the subject of many inquiries.

In a remarkable series of papers ([1], [2], [3]) Slepian, Landau, and Pollak demonstrated the key role of the prolate spheroidal wave functions in understanding this problem. The prolate spheroidal wave functions were shown to be the eigenfunctions of the operator that timelimits, then bandlimits, and then again timelimits a function. This interpretation of the prolate spheroidal wave functions led to a very exact picture of the trade-off involved in the time and bandlimiting of functions, including a generalization of the Heisenberg uncertainty principle.

In a recent series of papers ([4], [5], [6], [7]) the notion of "time" and "bandlimiting" was explored for more general "Fourier-type" situations, in particular for expansions in terms of orthogonal polynomials. The operator that "time", then "band", and then "time limits" a function was studied and, for some special cases, an efficient means of determining the eigenfunctions of this operator was given (analogous to the method employed in [1]). The purpose of this note is to take the eigenfunctions generated in this new situation and show that they too enable us to draw an accurate picture of the extent to which functions can be "time" and bandlimited". One should note that the techniques employed in [2] generalize readily to this new situation and thus the real work has already been done. We further remark that many of the properties we will be discussing were shown in [8] for the case of expansions in Legendre polynomials.

2. Background. Let f be a square integrable function with Fourier transform $\hat{f}$. Let A be the operator that timelimits f and let B be the operator that bandlimits f, i.e.

$$Af = f \cdot \chi_{\mathscr{A}}, \qquad B\hat{f} = \hat{f} \cdot \chi_{\mathscr{B}}.$$

Let F and F^{-1} denote the operations of Fourier transform and inverse Fourier transform:

$$F(f) = \hat{f}, \qquad F^{-1}(\hat{f}) = f.$$

The (self-adjoint) operator that time-band-timelimits a function is given by

$$AF^{-1}BFA = E^*E, \quad \text{where } E = BFA.$$

*Received by the editors September 15, 1983, and in revised form April 5, 1984. This work was supported in part by the National Science Foundation under grant MCS-8302526.

†Department of Mathematics, Drexel University, Philadelphia, Pennsylvania 19104.

The operator E^*E is a finite convolution integral operator. For the case when $\mathscr{A}=[-T,T]$ and $\mathscr{B}=[-W,W]$,

$$E^*Ef(x)=\int_{\mathscr{A}} \frac{\sin W(x-y)}{x-y} f(y)\,dy, \qquad x\in\mathscr{A}$$

and the eigenfunctions of E^*E are the prolate spheroidal wave functions [1].

Extensions of this work to more general "Fourier-type" situations were discussed in [4], [5], [6], [7]. In particular we will consider complete orthogonal families of polynomials $\{p_i(x)|i=0,1,2,\cdots\}$ with continuous nonnegative weight functions $w(x)$ on set $\mathscr{C}$. Thus for f square integrable (with respect to $w(x)$, $\mathscr{C}$) we can write

$$f(x)=\sum_{i=0}^{\infty} \hat{f}(i)p_i(x), \qquad \hat{f}(i)=\langle f(x),p_i(x)\rangle_{w(x),\mathscr{C}}$$

where the $p_i(x)$'s are suitably normalized and

$$\langle g,h\rangle_{w(x),\mathscr{C}}=\int_{\mathscr{C}} g(x)h(x)w(x)\,dx.$$

Our analogues of time and bandlimiting will be given as follows:

$$Af=f\cdot\chi_{\mathscr{A}}$$

$$B\hat{f}(i)=\begin{cases}\hat{f}(i), & i=0,1,\cdots,L,\\ 0, & i>L\end{cases}$$

where $\mathscr{A}\subset\mathscr{C}$ has positive measure strictly less than the measure of $\mathscr{C}$, i.e.

$$0<\int_{\mathscr{A}} w(x)\,dx<\int_{\mathscr{C}} w(x)\,dx. \tag{2.0}$$

Thus

$$E^*Ef(x)=\int_{\mathscr{A}} K_L(x,y)f(y)w(y)\,dy, \qquad x\in\mathscr{A}$$

where

$$K_L(x,y)=\sum_{i=0}^{L} p_i(x)p_i(y).$$

We note that in addition to the time-band-timelimiting operator E^*E, one can equally well study the band-time-bandlimiting operator EE^*. This operator is given by the $(L+1)\times(L+1)$ matrix G with entries

$$G_{ij}=\langle p_i(x),p_j(x)\rangle_{w(x),\mathscr{A}}=\int_{\mathscr{A}} p_i(x)p_j(x)w(x)\,dx, \qquad 0\leqq i\leqq j\leqq L.$$

The duality between the operators E^*E and EE^* is given in Lemma 2.1 below and will be exploited shortly.

LEMMA 2.1. *If f is an eigenfunction of $E^*E=AF^{-1}BFA$ with eigenvalue $\lambda\neq 0$, then BFf is an eigenvector of $EE^*=BFAF^{-1}B$ with eigenvalue λ. Similarly if $\hat{f}$ is an eigenvector of EE^* with eigenvalue $\lambda\neq 0$, then $AF^{-1}\hat{f}$ is an eigenfunction of E^*E with eigenvalue λ.*

For a proof of Lemma 2.1, see [4].

We also remark that by taking $\mathscr{A} \subset \mathscr{C}$ as in (2.0), we are assured of the existence of a complete orthogonal family of polynomials on $\mathscr{A}$ with respect to $w(x)$. The completeness of polynomials for $L^2_{w(x),\mathscr{A}}$ follows immediately from their completeness in $L^2_{w(x),\mathscr{C}}$. The linear independence of $\{1, x, x^2, \cdots\}$ in $\mathscr{A}$ follows readily since $\mathscr{A}$ has positive measure and any polynomial $p(x)$ has only a finite number of zeros.

3. The eigenfunctions of E^*E: polynomial analogues of the prolate spheroidal wave functions. It follows from Lemma 2.1 that for any expansion in terms of a complete orthogonal polynomial family and set $\mathscr{A} \subset \mathscr{C}$ ($\mathscr{A}$ as in (2.0)) the eigenfunctions of E^*E can be obtained by determining the eigenvectors of matrix G. In particular if $\mathbf{c} = (c_0, c_1, \cdots, c_L)$ is an eigenvector of G with eigenvalue $\lambda \neq 0$, then setting $\phi(x) = \sum_{i=0}^{L} c_i p_i(x)$, we have $A\phi(x) = \phi^{\mathscr{A}}(x)$ is an eigenfunction of E^*E (we will use $f^{\mathscr{A}}(x)$ to indicate Af). Note that $\phi(x)$ is "bandlimited" ($\phi(x)$ is a polynomial of degree $\leqq L$) and that $\phi^{\mathscr{A}}(x)$ is "timelimited". The $\phi(x)$'s are our polynomial analogues of the prolate spheroidal wave functions. Many of the properties of the $\phi(x)$'s listed below (such as their double orthogonality on $\mathscr{A}$ and $\mathscr{C}$, Lemma 3.1), are readily seen to be analogous to the properties of the prolate spheroidal wave functions (see [1]).

From Lemma 2.1 it follows that E^*E has at most $(L+1)$ linearly independent eigenfunctions corresponding to nonzero eigenvalues. If we denote these eigenfunctions (before chopping to $\mathscr{A}$) by $\phi_0(x), \phi_1(x), \cdots, \phi_k(x)$ with corresponding eigenvalues $\lambda_0 \geqq \lambda_1 \geqq \cdots \geqq \lambda_k$, then we note that in fact $k = L$. This follows since every function f, square integrable with respect to $\langle \cdot, \cdot \rangle_{w(x),\mathscr{A}}$, can be rewritten as

$$f^{\mathscr{A}}(x) = \sum_{i=0}^{k} a_i \phi_i^{\mathscr{A}}(x) + h^{\mathscr{A}}(x), \quad \text{where}$$

$$E^*Eh^{\mathscr{A}}(x) = 0 \quad \text{and} \quad \langle h^{\mathscr{A}}(x), \phi_i(x) \rangle_{w(x),\mathscr{A}} = 0, \qquad i = 0, 1, \cdots, k.$$

Thus for all $x \in \mathscr{A}$ we have

$$E^*Eh^{\mathscr{A}}(x) = \int_{\mathscr{A}} h(y) K_L(x,y) w(y)\, dy = \sum_{i=0}^{L} p_i(x) \langle h(y), p_i(y) \rangle_{w(x),\mathscr{A}} = 0.$$

and conclude that $\langle h(y), p_i(y) \rangle_{w(x),\mathscr{A}} = 0$. Therefore for any polynomial $q(x)$ of degree $\leqq L$, $\langle h(x), q(x) \rangle_{w(x),\mathscr{A}} = 0$. Since there is a complete orthogonal family of polynomials $q_i(x)$ ($i = 0, 1, 2, \cdots$, degree $q_i(x) = i$) with respect to $\langle \cdot, \cdot \rangle_{w(x),\mathscr{A}}$, we have that for $i = 0, 1, \cdots, L$, $q_i(x)$ can be written as a sum of the $\phi_j(x)$'s, $j = 0, 1, \cdots, k$. Thus $k = L$ and we see that $\{q_i(x) | i = L+1, L+2, \cdots\}$ form a basis for $H = \{h(x) | E^*Eh(x) = 0\}$.

The lemmas below summarize a number of properties of the ϕ_i's that will be used in the next section.

LEMMA 3.1. *Let $\{p_i(x)\}$ be a complete orthogonal family of polynomials ($i = 0, 1, 2, \cdots$ degree $p_i(x) = i$) with respect to the nonnegative continuous weight function $w(x)$ on $\mathscr{C}$. Let $\mathscr{A} \subset \mathscr{C}$ be as in* (2.0). *Then the operator E^*E has $L+1$ linearly independent eigenfunctions $\phi_0(x), \phi_1(x), \cdots, \phi_L(x)$ corresponding to the nonzero eigenvalues $1 > \lambda_0 \geqq \lambda_1 \geqq \cdots \geqq \lambda_L > 0$. Without loss of generality the ϕ_i's can be normalized so that*

(a) $\int_{\mathscr{A}} K_L(x,y) \phi_i(y) w(y)\, dy = \lambda_i \phi_i(x)$ *for all* $x \in \mathscr{C}$.

(b) $\langle \phi_i(x), \phi_j(x) \rangle_{w(x),\mathscr{C}} = \delta_{ij}$.

(c) $\langle \phi_i(x), \phi_j(x) \rangle_{w(x),\mathscr{A}} = \lambda_i \delta_{ij}$.

(d) *The* ϕ_i's *comprise a complete orthonormal set for the space of bandlimited functions.*

(e) *The* $\phi_i^{\mathscr{A}}$'s *together with the polynomials* $q_{L+1}, q_{L+2}, \cdots$ *where* $\{q_i(x)\}$ *is the orthogonal polynomial family for* $\langle \cdot, \cdot \rangle_{w(x),\mathscr{A}}$ *form a complete orthogonal set for the space of square integrable timelimited functions.*

Proof. (a) Corresponding to each ϕ_i is an eigenvector $\mathbf{c} = (c_0, \cdots, c_L)$ of EE^*:

$$BFAF^{-1}B\mathbf{c} = \lambda_i \mathbf{c}.$$

Applying F^{-1} on both sides of the above yields (a).

(b) *and* (c) That the ϕ_i's can be chosen orthogonal on $\mathscr{A}$ follows since E^*E is self-adjoint. From (a) and the fact that

$$\int_{\mathscr{C}} K_L(x,y) K_L(x,z) w(x)\, dx = K_L(y,z),$$

we have that

$$\begin{aligned}\langle \phi_i, \phi_j \rangle_{w(x),\mathscr{C}} &= \frac{1}{\lambda_i \lambda_j} \int_{\mathscr{C}} \left(\int_{\mathscr{A}} \phi_i(y) K_L(x,y) w(y)\, dy \right) \\ &\quad \cdot \left(\int_{\mathscr{A}} \phi_j(z) K_L(x,z) w(z)\, dz \right) w(x)\, dx \\ &= \frac{1}{\lambda_i \lambda_j} \int_{\mathscr{A}} \int_{\mathscr{A}} \phi_i(y) \phi_j(z) w(y) w(z) \left[\int_{\mathscr{C}} K_L(x,y) K_L(x,z) w(x)\, dx \right] dy\, dz \\ &= \frac{1}{\lambda_i} \int_{\mathscr{A}} \phi_i(y) \phi_j(y) w(y)\, dy = \frac{1}{\lambda_i} \langle \phi_i, \phi_j \rangle_{w(x),\mathscr{A}}.\end{aligned}$$

Thus orthogonality on $\mathscr{A}$ implies orthogonality on $\mathscr{C}$ and without loss of generality the ϕ_i's can be chosen to be orthonormal on $\mathscr{C}$ and (b) and (c) hold. Furthermore taking $i = j$ in the above shows $0 < \lambda_i \leqq 1$. That $\lambda_i < 1$ follows from Lemma 3.2 below.

(d) *and* (e) are clear from the earlier discussion in this section.

We will also be interested in looking at our operators on the set $\sim\mathscr{A}$ where $\sim\mathscr{A}$ is the complement of $\mathscr{A}$ in $\mathscr{C}$ ($\sim\mathscr{A} = \mathscr{C} - \mathscr{A}$). Using the notation

$$\sim A f(x) = f(x) \cdot \chi_{\sim\mathscr{A}}(x) = f^{\sim\mathscr{A}}(x),$$

we see that

LEMMA 3.2. (a) $\langle \sim A\phi_i, \sim A\phi_i \rangle_{w(x),\mathscr{C}} = 1 - \lambda_i$.

(b) $[\sim AF^{-1}BF \sim A]\phi_i(x) = (1 - \lambda_i)\phi_i(x)$.

Proof. (a) is immediate.

(b) follows since $F^{-1}BF\phi_i = \phi_i$ and thus $F^{-1}BF(A + \sim A)\phi_i = \phi_i$. Therefore, $F^{-1}BFA\phi_i + F^{-1}BF \sim A\phi_i = \lambda_i \phi_i + (1 - \lambda_i)\phi_i$.

We will also need:

LEMMA 3.3. *For* $h(x)$ *such that* $E^*Eh(x) = 0$ $(x \in \mathscr{A})$ *we have in fact* $F^{-1}BFAh(x) = 0$ *for all* $x \in \mathscr{C}$.

Proof.

$$F^{-1}BFAh(x) = \sum_{i=0}^{L} p_i(x) \int_{\mathscr{A}} h(y) p_i(y) w(y)\, dy = 0$$

for all $x \in \mathscr{C}$ since

$$\int_{\mathscr{A}} h(y)\, p_i(y)\, w(y)\, dy = 0 \quad \text{for } i = 0, 1, \cdots, L.$$

4. The uncertainty principle. The Heisenberg uncertainty principle measures how closely a function f and its Fourier transform ($\hat{f} = \int_{-\infty}^{\infty} f(s) e^{-i\delta s}\, ds$) can be simultaneously concentrated about a point. In particular, if

$$\mathscr{T}^2 = \frac{\int (x - x_0)^2 |f(x)|^2\, dx}{\int |f(x)|^2\, dx} \quad \text{and} \quad \Omega^2 = \frac{\int (\delta - \delta_0)^2 |\hat{f}(\delta)|^2\, d\delta}{\int |\hat{f}(\delta)|^2\, d\delta},$$

then for any x_0, δ_0 we have $\Omega\mathscr{T} \geqq \frac{1}{2}$ and thus Ω and $\mathscr{T}$ cannot both be small [9].

In many situations, however, one is more interested in the extent to which a function can be concentrated on some subset of the real line (e.g. an interval). Thus in [2] Landau and Pollak studied the quantities

$$\alpha^2 = \frac{\|Af\|_2^2}{\|f\|_2^2} \quad \text{and} \quad \beta^2 = \frac{\|B\hat{f}\|_2^2}{\|\hat{f}\|_2^2}.$$

Using the fact that the prolate spheroidal wave functions are the eigenfunctions of E^*E, they were able to determine all possible values of (α, β) for f square integrable. Thus, for example, to determine how nearly "timelimited" a "bandlimited" function can be, simply maximize α for $\beta = 1$.

What we will do here is to consider the quantities α and β for the case of orthogonal polynomial expansions. The theorem below will follow from minor modifications of the arguments given in [2].

THEOREM 4.1. *Let* $\{p_i(x)\}$, $i = 0, 1, 2, \cdots$, *be a complete orthogonal polynomial family* (*degree* $p_i(x) = i$) *with respect to the inner product* $\langle f, g \rangle_{w(x), \mathscr{C}}$ *where* $w(x)$ *is nonnegative and continuous. Let* $\mathscr{A} \subset \mathscr{C}$ *be as in* (2.0). *Then there is a square integrable function*

$$f(x) = \sum_{i=0}^{\infty} \hat{f}(i)\, p_i(x)$$

such that

$$\alpha^2 = \frac{\|Af\|_2^2}{\|f\|_2^2} \quad \textit{and} \quad \beta^2 = \frac{\|B\hat{f}\|_2^2}{\|\hat{f}\|_2^2} = \frac{\sum_{i=0}^{L} \hat{f}^2(i)}{\sum_{i=0}^{\infty} \hat{f}^2(i)}$$

if and only if

(a) *for* $\beta^2 = 1$, $\lambda_L \leqq \alpha^2 \leqq \lambda_0$,

(b) *for* $\alpha^2 = 1$, $0 \leqq \beta^2 \leqq \lambda_0$,

(c) *for* $\alpha^2 = 0$, $0 \leqq \beta^2 \leqq 1 - \lambda_L$,

(d) *for* $\beta^2 = 0$, $0 \leqq \alpha^2 \leqq 1$,

(e) *for* $\lambda_L \leqq \alpha^2 \leqq \lambda_0$, $0 \leqq \beta^2 \leqq 1$,

(f) *for* $\lambda_0 \leqq \alpha^2 \leqq 1$, $\cos^{-1}\alpha + \cos^{-1}\beta \geqq \cos^{-1}\sqrt{\lambda_0}$ $(\beta \geqq 0)$,

(g) *for* $0 \leqq \alpha^2 \leqq \lambda_L$, $\cos^{-1}(1-\alpha) + \cos^{-1}\beta \geqq \cos^{-1}\sqrt{1-\lambda_L}$ $(\beta \geqq 0)$.

Before proving Theorem 4.1 several comments are in order. If we look at all possible pairs (α, β) in the square $0 \leqq \alpha^2 \leqq 1$, $0 \leqq \beta^2 \leqq 1$, then (a)–(d) tell us the attainable boundary values, indicated by the darkened lines in Fig. 1 and (e) is handled by the

shaded area

The areas marked by

follow readily from (f) and (g). Thus it is only in the two upper corners that the picture becomes complicated. The similarity of this result with the result for the standard Fourier transform in [2] is readily apparent. The only real change here is the upper left-hand corner.

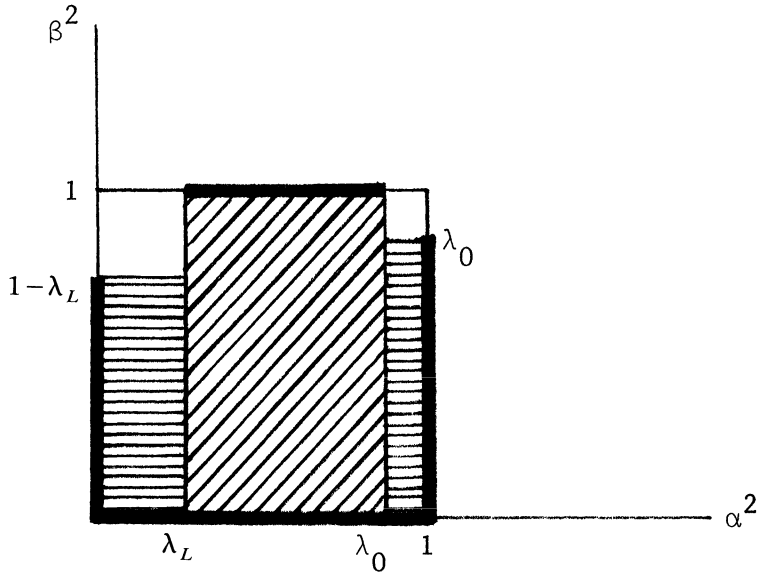


FIG. 1

Some information about the geometry of the space $\mathscr{W}+\mathscr{T}$ where $\mathscr{W}=\{\text{bandlimited functions}\}$ and $\mathscr{T}=\{\text{timelimited functions}\}$ will be needed for the proof of Theorem 4.1. These results were obtained for the standard Fourier transform in [2] and the proofs there apply readily to our more general situation and will not be repeated here. A brief outline of the results will be given below.

Define the angle θ between two functions f and g ($\|f\|, \|g\| > 0$) as

$$\theta(f,g) = \cos^{-1} \frac{\langle f,g \rangle}{\|f\| \, \|g\|}.$$

Note that for $f \in \mathscr{W}$ and $g \in \mathscr{T}$, $\theta(f,g) > 0$ since $\mathscr{W} \cap \mathscr{T}$ contains only $f = 0$. Landau and Pollak show that there is a least angle between $\mathscr{W}$ and $\mathscr{T}$;

THEOREM 4.2. *There exists a least angle between $\mathscr{W}$ and $\mathscr{T}$. This angle equals* $\cos^{-1}\sqrt{\lambda_0}$ *and is assumed by* $\phi_0 \in \mathscr{W}$ *and* $D\phi_0 \in \mathscr{T}$, i.e.,

$$\min_{\substack{f \in \mathscr{W} \\ g \in \mathscr{T}}} \phi(f,g) = \cos^{-1}\sqrt{\lambda_0}.$$

Landau and Pollak also give the following lemma which extends directly to the situation we are studying:

LEMMA 4.3. *$\mathscr{W}+\mathscr{T}$ is closed.*

Proof of Theorem 4.1. (a) $\beta=1$: Thus f is "bandlimited," $f(x)=\sum_{i=0}^{L} c_i\phi_i(x)$ and without loss of generality $\|f\|=\|\hat{f}\|=1$. Thus

$$Af=\sum_{i=0}^{L} c_i\phi_i(x), \qquad \alpha^2=\sum_{i=0}^{L} c_i^2\lambda_i \quad \text{where} \quad \sum_{i=0}^{L} c_i^2=1.$$

Clearly α^2 is maximized for $c_0=1$ ($f=\phi_0$) and minimized for $c_L=1$ ($f=\phi_L$). Thus $\alpha^2\leqq\lambda_0$ and $\lambda_L\leqq\alpha^2$. To see that we can obtain all values of α^2 between λ_0 and λ_L, consider $g=a\phi_0+b\phi_L$ where $a^2+b^2=1$. For g we have $\beta^2=1$ and $\alpha^2=a^2\lambda_0+b^2\lambda_L$. Thus as a, b range from 0 to 1 all values of α^2 between λ_0 and λ_L are attained.

(b) $\alpha=1$. Thus f is "timelimited". If $f=\phi_0^{\mathscr{A}}$ then $\alpha^2=1$ and since $FBF^{-1}A\,\phi_0^{\mathscr{A}}=\lambda_0\phi_0$, $\beta^2=\lambda_0$. If $f=h^{\mathscr{A}}(x)$ ($E^*Eh(x)=0$) then $\alpha^2=1$, $\beta^2=0$. That all values of β^2 between 0 and λ_0 are attained can be seen by taking $g(x)=a\phi_0^{\mathscr{A}}(x)+bh^{\mathscr{A}}(x)$ with $0\leqq a,\ b\leqq 1$. Thus $\alpha^2=1$ and $F^{-1}BFAg(x)=a\lambda_0\phi_0(x)$ which gives $\beta^2=a^2\lambda_0^2/(a^2\lambda_0+b^2\mu)$ where $\mu=\|h^{\mathscr{A}}(x)\|_2^2$. This produces all intermediate values.

That β^2 cannot be greater than λ_0 follows from Theorem 4.2 (see also the proof of part (f)).

(c) $\alpha=0$. This is the opposite of $\alpha=1$ in that we now have $f^{\sim\mathscr{A}}(x)=f(x)$. Taking $f(x)=\sum_{i=0}^{L} c_i\phi_i^{\sim\mathscr{A}}(x)+h_2^{\sim\mathscr{A}}(x)$ where $h_2^{\sim\mathscr{A}}(x)$ satisfies $\sim AF^{-1}BF\sim Ah_2^{\sim\mathscr{A}}(x)=0$, we have $F^{-1}BF\sim Af(x)=\sum_{i=0}^{L} c_i(1-\lambda_i)\phi_i(x)$ and reasoning analogous to that in (b) yields $0\leqq\beta^2\leqq 1-\lambda_L$.

(d) $\beta=0$. Choose $h_1^{\mathscr{A}}(x)$ so that $E^*Eh_1^{\mathscr{A}}(x)=0$ and let $h_2^{\sim\mathscr{A}}(x)$ be as in part (c). Thus for $h_1^{\mathscr{A}}(x)$, $\beta=0$ and $\alpha=1$ and for $h_2^{\sim\mathscr{A}}(x)$, $\beta=0$ and $\alpha=0$. That all values of α between 0 and 1 can be attained is seen by taking $g(x)=a_1h_1^{\mathscr{A}}(x)+a_2h_2^{\sim\mathscr{A}}(x)$ where $0<a_1,\ a_2<1$.

(e) $\lambda_L\leqq\alpha^2\leqq\lambda_0$. Fix some α, $\lambda_L\leqq\alpha^2\leqq\lambda_0$. Thus we can find f and g such that:

$$\|f\|=1, \quad \|f^{\mathscr{A}}\|=\alpha, \quad \beta=0, \quad f=a_1h_1^{\mathscr{A}}+a_2h_2^{\sim\mathscr{A}} \quad \text{(see (d))}$$

and

$$\|g\|=1, \quad \|g^{\mathscr{A}}\|=\alpha, \quad \beta=1, \quad g=b_1\phi_0+b_2\phi_L \quad \text{(see (a))}.$$

Since $\langle f,g\rangle_{w(x),\mathscr{C}}=\langle f,g\rangle_{w(x),\mathscr{A}}=0$, letting $t(x)=c_1f(x)+c_2g(x)$, $c_1^2+c_2^2=1$, we can attain all values of β between 0 and 1.

(f) $\lambda_0\leqq\alpha^2\leqq 1$. We begin by showing that if $\|f\|=1$ and $\|Df\|=\alpha$, then $\beta\leqq\cos(\cos^{-1}\sqrt{\lambda_0}-\cos^{-1}\alpha)$ with equality attained for

$$f=p\phi_0+qD\phi_0 \quad \text{with } p=\sqrt{\frac{1-\alpha^2}{1-\lambda_0}} \quad \text{and} \quad q=\frac{\alpha}{\sqrt{\lambda_0}}-\sqrt{\frac{1-\alpha^2}{1-\lambda_0}}\,.$$

This result follows exactly from the argument in [2] and we will just briefly outline the steps here. Since $\mathscr{W}+\mathscr{T}$ is closed for any f with $\|f\|=1$ and $\|f^{\mathscr{A}}\|=\alpha$, we can find g orthogonal to $f^{\mathscr{A}}$ and $f^{\mathscr{B}}=F^{-1}BFf$ such that

$$f=\lambda f^{\mathscr{A}}+\mu f^{\mathscr{B}}+g. \tag{4.4}$$

By taking inner products of (4.4) with $f, f^{\mathscr{A}}, f^{\mathscr{B}}$, and g and eliminating (g,f), λ, and μ from the resulting equations, we get (for $\alpha\beta \neq 0$):

$$\beta^2 - 2\langle f^{\mathscr{A}}, f^{\mathscr{B}}\rangle = -\alpha^2 + 1 - \frac{\langle f^{\mathscr{A}}, f^{\mathscr{B}}\rangle^2}{\alpha^2\beta^2} - \|g\|^2\left(1 - \frac{\langle f^{\mathscr{A}}, f^{\mathscr{B}}\rangle^2}{\alpha^2\beta^2}\right).$$

Setting

$$\cos\theta = \frac{\langle f^{\mathscr{A}}, f^{\mathscr{B}}\rangle}{\|f^{\mathscr{A}}\|\|f^{\mathscr{B}}\|},$$

one obtains

$$(\beta - \alpha\cos\theta)^2 \leqq (1-\alpha^2)\sin^2\theta$$

with equality if and only if $g = 0$. Therefore

$$\beta \leqq \cos(\theta - \cos^{-1}\alpha)$$

and since by Theorem 4.2, $\theta \geqq \cos^{-1}\sqrt{\lambda_0}$,

$$\cos^{-1}\alpha + \cos^{-1}\beta \geqq \cos^{-1}\sqrt{\lambda_0}. \tag{4.5}$$

Equality in (4.5) is attained by

$$f(x) = p\phi_0(x) + q\phi_0^{\mathscr{A}}(x) \tag{4.6}$$

with

$$p = \sqrt{\frac{1-\alpha^2}{1-\lambda_0}} \quad \text{and} \quad q = \frac{\alpha}{\sqrt{\lambda_0}} - \sqrt{\frac{1-\alpha^2}{1-\lambda_0}}.$$

To see that in fact $0 \leqq \beta \leqq \cos(\cos^{-1}\sqrt{\lambda_0} - \cos^{-1}\alpha)$ simply consider

$$cf(x) + d\frac{g(x)}{\|g(x)\|}$$

where $g(x)$ is as in the proof of (d) ($g(x) = a_1 h_1^{\mathscr{A}}(x) + a_2 h_2^{\sim\mathscr{A}}(x)$) and $f(x)$ is as in 4.6 and $c^2 + d^2 = 1$.

(g) $0 \leqq \alpha^2 \leqq \lambda_L$: Consider all $f \in \mathscr{L}^2$ with $\|f\| = 1$, $\|f^{\mathscr{A}}\| = \alpha$. We first determine the maximum β. This is really equivalent to part (f) with $\alpha' = \|f^{\sim\mathscr{A}}\| = 1 - \alpha$ and $\beta' = \beta$. Thus

$$0 \leqq \beta' = \beta \leqq \cos\left(\cos^{-1}\sqrt{1-\lambda_L} - \cos^{-1}(\alpha')\right)$$

completing the proof.

We note that when $\beta = 1$, $\alpha^2 \leqq \lambda_0$ and $\alpha^2 = \lambda_0$ for $f(x) = \phi_0(x)$. Thus $\phi_0(x)$ is the polynomial of degree $\leqq L$ that has the greatest proportion of its norm in $\mathscr{A}$. Similarly $\phi_L(x)$ represents the polynomial of degree $\leqq L$ that has the least proportion of its norm in $\mathscr{C}$. For some graphs of these functions for the case of the Legendre polynomials ($w(x) = 1$, $\mathscr{C} = [-1, 1]$) see [8].

5. Conclusion. Recall that the values λ_0 and λ_L needed in Theorem 4.1 are respectively the largest and smallest eigenvalues of the matrix G of §2. Thus if L is not too large (and the entries of G are available) the work involved in the computation of λ_0 and λ_L is not excessive. In certain special cases (notably the classical orthogonal polynomials where $\mathscr{C}=(\delta,\varepsilon)$ and $\mathscr{A}=(\delta',\varepsilon)$, $\delta<\delta'$), there is a relatively simple means of obtaining the eigenvectors ([4], [5]).

It should also be noted that while the prolate spheroidal wave functions depend only on the product $c=WT$, the eigenfunctions in the polynomial case depend on both L and $\mathscr{A}$. Thus in [2], the authors are able to write a Heisenberg-like bound:

$$WT \geqq \phi(\alpha,\beta), \quad \text{where } \mathscr{A}=[-W,W], \quad \mathscr{B}=[-T,T]$$

for some explicit function ϕ. In the polynomial case we have been unable to do this.

REFERENCES

[1] D. SLEPIAN AND H. O. POLLAK, *Prolate spheroidal wave functions, Fourier analysis and uncertainty*: I, Bell System Tech. J., 40 (1961), pp. 43–64.

[2] H. J. LANDAU AND H. O. POLLAK, *Prolate spheroidal wave functions, Fourier analysis and uncertainty*: II, Bell System Tech. J., 40 (1961), pp. 65–84.

[3] _____, *Prolate spheroidal wave functions, Fourier analysis and uncertainty*: III, Bell System Tech. J., 41 (1962), pp. 1295–1336.

[4] F. A. GRÜNBAUM, L. LONGHI AND M. PERLSTADT, *Differential operators commuting with finite convolution integral operators: Some nonabelian examples*, SIAM J. Appl. Math., 42 (1982), pp. 941–955.

[5] F. A. GRÜNBAUM, *A new property of a reproducing kernel for classical orthogonal polynomials*, J. Math. Anal. Appl., 95 (1983), pp. 491–500.

[6] M. PERLSTADT, *Chopped orthogonal polynomial expansions—Some discrete cases*, SIAM J. Alg. Disc. Meth., 4 (1983), pp. 94–100.

[7] _____, *A property of orthogonal polynomial families with polynomial duals*, this Journal, 15 (1984), pp. 1042–1054.

[8] E. N. GILBERT AND D. SLEPIAN, *Doubly concentrated orthogonal polynomials*, this Journal, 8 (1977), pp. 290–319.

[9] H. DYM AND H. P. MCKEAN, *Fourier Series and Integrals*, Academic Press, New York, 1972.

SIAM J. MATH. ANAL.
Vol. 17, No. 2, March 1986

SYMMETRY AND STABILITY IN TAYLOR-COUETTE FLOW*

MARTIN GOLUBITSKY† AND IAN STEWART‡

Abstract. We study the flow of a fluid between concentric rotating cylinders (the Taylor problem) by exploiting the symmetries of the system. The Navier–Stokes equations, linearized about Couette flow, possess two zero and four purely imaginary eigenvalues at a suitable value of the speed of rotation of the outer cylinder. There is thus a reduced bifurcation equation on a six-dimensional space which can be shown to commute with an action of the symmetry group $O(2)\times SO(2)$. We use the group structure to analyze this bifurcation equation in the simplest (nondegenerate) case and to compute the stabilities of solutions. In particular, when the outer cylinder is counterrotated we can obtain transitions which seem to agree with recent experiments of Andereck, Liu, and Swinney [1984]. It is also possible to obtain the "main sequence" in this model. This sequence is normally observed in experiments when the outer cylinder is held fixed.

Introduction. The flow of a fluid between concentric rotating cylinders, or *Taylor–Couette flow*, is known to exhibit a variety of types of behavior, the most celebrated being *Taylor vortices* (Taylor [1923]). The problem has been studied by a large number of authors: a recent survey is that of DiPrima and Swinney [1981]. The experimental apparatus has circular symmetry, and the standard mathematical idealization (periodic boundary conditions at the ends of the cylinder) introduces a further symmetry. As a result the Navier–Stokes equations for this problem are covariant with respect to the action of a symmetry group $O(2)\times SO(2)$. It has become clear that the symmetries inherent in bifurcating systems have a strong influence on their behavior. In this paper we study a series of bifurcations that occur in Taylor–Couette flow placing emphasis on the role of symmetry. (Schecter [1976] and Chossat and Iooss [1984] have also studied the problem from this viewpoint, and we discuss the relations between our work and theirs below.)

DiPrima and Grannick [1971] have found that when the outer cylinder is rotated in a direction opposite to that of the inner cylinder, the Navier–Stokes equations, linearized about Couette flow, possess six eigenvalues on the imaginary axis. It follows that aspects of the dynamics can be reduced (either by Lyapunov–Schmidt or center manifold reduction) to a vector field on $\mathbf{R}^6$; furthermore, this vector field commutes with an action of $O(2)\times SO(2)$. Moreover, as we explain in §7, recent experimental results due to Andereck, Liu, and Swinney [1984] seem to confirm the existence of the six-dimensional kernel.

We point out in particular that the six-dimensional kernel is a codimension one phenomenon, and hence it is not surprising that it should be possible to find it by varying only one parameter. Indeed, this degeneracy should occur relatively often in various circumstances, and so deserves detailed analysis.

We study the general class of bifurcation problems on $\mathbf{R}^6$ having this $O(2)\times SO(2)$ symmetry. We derive the general form possible for the vector field, and by classifying

* Received by the editors November 15, 1984, and in revised form February 8, 1985.

† Department of Mathematics, University of Houston-University Park, Houston, Texas 77004. The research of this author was supported in part by the National Science Foundation under grant MCS-8101580, and by grant NAG 2-279 from NASA-Ames.

‡ Department of Mathematics, University of Houston-University Park, Houston, Texas 77004 and Mathematics Institute, University of Warwick, Coventry, CV4 7AL, England. The research of this author was supported in part by grant NAG 2-279 from NASA-Ames.

the possible ways to break symmetry, obtain equations for the bifurcating branches (subject to certain nondegeneracy conditions). We also obtain the (linearized orbital) stabilities of these branches.

By introducing an additional parameter α we split the kernel $\mathbf{R}^6$ into two subspaces $\mathbf{R}^2$ and $\mathbf{R}^4$ corresponding to a steady-state and a periodic bifurcation respectively. Depending on the sign of α, one or other of these bifurcations occurs first.

By inspecting the symmetries of the physically observed solutions we may tentatively identify them with various branches: in particular the flows known as Taylor vortices, wavy vortices, twisted vortices, helices (or spirals) seem to correspond naturally to solution branches; and there is also a branch described by DiPrima and Grannick [1971] as the "nonaxisymmetric simple mode".

The experimental results of Andereck, Liu, and Swinney may be summarized as follows. In the weakly counterrotating case (that is, when the speed of the outer cylinder Ω_0 is slightly less than the critical speed Ω_0^* where the six-dimensional kernel appears) the following transition sequence is observed as Ω_i, the speed of the inner cylinder, is increased.

$$\text{Couette flow} \to \text{Taylor vortices} \to \text{wavy vortices} \to \cdots$$

where the final state obtained when the wavy vortices lose stability seems not to be one representable in the six-dimensional kernel. In the strongly counterrotating case (that is, when Ω_0 is slightly greater than Ω_0^*) the observed transition sequence is:

$$\text{Couette flow} \to \text{spiral cells} \to \text{wavy spiral cells}.$$

We shall show in §7 that it is possible to make a nondegenerate choice of vector fields on $\mathbf{R}^6$ having $O(2)\times SO(2)$ symmetry which produces the same transition sequences in the following sense. It is possible to determine constraints on the Taylor expansion of this vector field, given only by inequalities on coefficients in this Taylor expansion, so that the solutions corresponding to these states are (orbitally) asymptotically stable and lose stability in a way that should produce the desired transitions. Moreover, when these inequalities are satisfied, no other solutions are asymptotically stable.

We also show in §7 that it is possible to choose these constraints differently, so that the "main sequence" of transitions occurs, namely,

$$\text{Couette flow} \to \text{Taylor vortices} \to \text{wavy vortices}$$
$$\to \cdot\,\text{modulated wavy vortices} \to \cdots.$$

This transition sequence is usually observed when the outer cylinder is held stationary ($\Omega_0 = 0$). What we show is that it is possible for the "wavy vortex solutions" to lose stability to a torus bifurcation, where two Floquet exponents cross the imaginary axis. This tertiary bifurcation has never been demonstrated theoretically hitherto. At this point, however, we cannot prove that the branch of "modulated wavy vortices" is asymptotically stable, though we hope that the results of Scheurle and Marsden [1984] will provide the techniques required to carry out this computation. We do show, moreover, that no other solutions are asymptotically stable when these constraints hold. In particular, stable spiral cells should not occur in this experimental situation.

The paper is organized as follows. In §1 we describe some of the flows observed in the Taylor experiment and review the evidence for the existence of a six-dimensional kernel. In §2 we discuss the symmetries that act on the six-dimensional kernel, and in §3 we discuss the symmetries of the observed flows. In §4 (and the Appendix) we

discuss the reduction procedure and derive the exact form of the reduced mapping (or vector field) on the six-dimensional kernel prescribed by those symmetries. We classify the (conjugacy classes) of isotropy subgroups (which describe the type of symmetry-breaking that occurs at bifurcations). The heart of the paper is §5, where we analyze the branching equations and the stability of branches. In §6 we use these to obtain a list of the sixteen inequalities that must be imposed to ensure (what we mean by) nondegeneracy. Finally, in §7, we compare the six-dimensional model with experimental observations in both the counterrotating case and the case when the outer cylinder is held fixed.

1. The Taylor problem. By the term "Taylor problem" we mean the study of both the possible states of fluid flow between two rotating concentric cylinders, and the transitions between these states. The Taylor problem provides a beautiful example of a bifurcation problem with symmetry. In this paper we discuss how these symmetries affect the structure of the bifurcating solutions.

We denote the angular velocities of the inner and outer cylinders by Ω_i and Ω_0 respectively. To specify a direction, we assume that $\Omega_i \geqq 0$. In the standard experiments the outer cylinder is held fixed ($\Omega_0 = 0$) and the inner cylinder is speeded up in stages from $\Omega_i = 0$, at each stage allowing the flow to settle into a stable pattern; see Taylor [1923], Gollub and Swinney [1975]. Experiments have been performed in both the *corotating* case ($\Omega_0 > 0$), see Andereck, Dickman and Swinney [1983], and the *counter-rotating* case ($\Omega_0 < 0$), see Andereck, Liu, and Swinney [1984]. (These papers cite the earlier experimental work.) The experiments begin by rotating the outer cylinder at constant speed, and allowing the flow to stabilize; then the inner cylinder is speeded up as before. The experiments reveal a large number of fluid states, only some of which are understood on theoretical grounds. There can exist multiple steady states whose exploration requires different experimental procedures; see for example Coles [1965], Benjamin [1978a, b], Benjamin and Mullin [1982]. In our discussion we shall assume a fixed (but unspecified) value of Ω_0, and treat Ω_i (or the corresponding Reynolds number) as a bifurcation parameter. Our main concern will be with the series of bifurcations that occurs as Ω_i is increased steadily. We mention this because many numerical computations fix the ratio Ω_0/Ω_i, and hence do not correspond directly to the usual experimental procedure—a fact that, in the presence of multiple states, raises some problems of interpretation.

In the standard experiments, with $\Omega_0 = 0$, the first transition is from Couette (laminar) flow to (Taylor) *vortices*. Both flows are time-independent. This transition was first described, in terms of a steady state bifurcation, by Davey [1962]. He showed that as Ω_i is increased, the Navier–Stokes equations linearized about Couette flow have a double zero eigenvalue at the first bifurcation. At this eigenvalue Couette flow loses stability, and a branch of vortex solutions bifurcates. Davey's observations have been reproduced by several authors in different contexts, cf. the survey by DiPrima and Swinney [1981, §6.3]. Note that the appearance of a double zero eigenvalue might be surprising were it not for the existence of symmetries (which can couple eigenvalues together and force a degeneracy).

Again, in the standard experiments with $\Omega_0 = 0$, a second transition is observed, in which vortices lose stability to a time-periodic state known as *wavy vortices*. Presumably this transition takes place by way of a Hopf type bifurcation in which several eigenvalues (governing the stability of vortices) cross the imaginary axis as Ω_i is increased. However, this presumption has never been established directly. What has been shown (in Davey, DiPrima, and Stuart [1968]) is that along the Couette branch of solutions

several eigenvalues of the linearized Navier–Stokes equations cross the imaginary axis as Ω_i is increased. In particular the next set of eigenvalues to cross the imaginary axis is a complex conjugate pair of purely imaginary eigenvalues, each of multiplicity two. Again, it would be surprising to see four eigenvalues crossing the imaginary axis simultaneously were it not for the symmetry. We note in passing (and amplify these remarks below) that the $O(2)$ symmetry which couples these four eigenvalues together forces the occurrence of two branches of time-periodic solutions bifurcating from the (unstable) main Couette branch: see Schecter [1976], and Golubitsky and Stewart [1985]. However, neither of these solutions can correspond to wavy vortex states, since their symmetries do not match those of wavy vortices. In fact, one of them has the symmetries of spiral cells (helices).

There are three additional facts which suggest that there might be a *relatively* simple local explanation for many of the observed states in the Taylor problem, at least in the counterrotating case $\Omega_0 < 0$. First, as observed in DiPrima and Grannick [1971], and Krueger, Gross, and DiPrima [1966], there is a critical speed of counterrotation $\Omega_0^* < 0$ such that, as Ω_i is increased, Couette flow loses linearized stability by having six eigenvalues cross the imaginary axis. These six eigenvalues are obtained by amalgamating the double zero eigenvalues and the complex conjugate pair of purely imaginary eigenvalues of multiplicity two, described above. Further, when Ω_0 is slightly less than Ω_0^*, the first bifurcation from Couette flow occurs when four eigenvalues (a complex conjugate pair each of multiplicity two) cross the imaginary axis; and there is a double zero eigenvalue at a higher value of Ω_i.

Second, in experiments in which Ω_0 is sufficiently negative, the primary bifurcation is not to the time-independent Taylor vortices, but to time-dependent spiral cells, see Andereck, Liu, and Swinney [1984].

Third, it is possible to produce a solution from the interaction of the four-dimensional center manifold (associated with the purely imaginary eigenvalues) and the two-dimensional center manifold (associated with the double zero eigenvalues) that has the same symmetry as wavy vortices. This suggests that it might be possible to prove the existence of a Hopf-type bifurcation from vortices to wavy vortices as a *secondary* bifurcation. This was observed by DiPrima and Sijbrand [1982] and again by Chossat and Iooss [1984].

Given these three facts, it would appear reasonable to study the Taylor problem in terms of perturbations of the degenerate case $\Omega_0 = \Omega_0^*$, using either a center manifold or a Lyapunov–Schmidt reduction from the Navier–Stokes equations. We call this degeneracy the *six-dimensional kernel* since the linearized equation has a kernel of dimension six and the reduced problem may therefore be posed on $\mathbf{R}^6$. The hope raised by the above facts is that one might be able to find a six-dimensional model which explains the observed prechaotic states and transitions in the counterrotating Taylor problem.

Let us consider the reduction in more detail. Rigorously, one can use the center manifold theorem to reduce the (infinite-dimensional) dynamics of the Navier–Stokes equations, near $\Omega_0 = \Omega_0^*$ and near Couette flow, to the study of some vector field g on a six-dimensional center manifold. Alternatively, one can focus only on time-independent and time-periodic solutions and use a reduction of the Lyapunov–Schmidt type to show the existence of a smooth (i.e. $\mathbf{C}^\infty$) mapping h: $\mathbf{R}^6 \to \mathbf{R}^6$ whose zeros are in one-to-one correspondence with the small-amplitude time-periodic (and time-independent) solutions of the Navier–Stokes equations. In either case, to study the dynamics $\dot{x} = g(x)$ on $\mathbf{R}^6$ or to solve $h(x) = 0$ in $\mathbf{R}^6$ would be a highly nontrivial task—were it not for the symmetries in the Taylor problem. Both reduction procedures can be performed so as to respect these symmetries. Therefore, g and h will commute with an action of

the symmetry group $O(2)\times S^1$ as we explain below. This places considerable restrictions on the form that g and h may take. When, as here, we are studying only steady and periodic states, it is sufficient to use the simpler Lyapunov–Schmidt reduction. This is our approach. For a complete study of the dynamics, the same restrictions on the form of g will be true *provided* a smooth center manifold exists. (It is plausible that the symmetry might imply this, but we have not attempted to address this issue here.)

In this paper we give an explicit representation for all smooth mappings that commute with this action of $O(2)\times S^1$. We use the symmetries to show how to solve the equation $h=0$ (in the Lyapunov–Schmidt interpretation), and to determine (in most instances) the signs of the eigenvalues of the 6×6 Jacobian matrix $dh\,|_{h=0}$. In particular we compute these eigenvalues for the solutions corresponding to wavy vortices.

In this respect our results resemble those of a recent paper of Chossat and Iooss [1984]. However, instead of working on the six-dimensional kernel, Chossat and Iooss track the bifurcations step by step using the symmetry in the primary bifurcation to analyze the possible types of symmetry-breaking at secondary bifurcations, in terms of the linearized eigenfunctions. The types of solution that they find can all be expressed as combinations of the six linearized eigenfunctions that make up the six-dimensional kernel; but no reduction to $\mathbf{R}^6$ is used explicitly. Thus, although the various pieces of the bifurcation diagram are studied, their overall arrangement (and consistency) is not.

In our approach group theory is used to provide a coherent framework that organizes the analysis and in particular the computation of stabilities, leading to more detailed results. In particular we confirm, in our setting, a conjecture made by Chossat and Iooss [1984] about *tertiary* bifurcation to modulated wavy vortices. We show that (with suitable parameter values) the branch of wavy vortices loses stability by a torus bifurcation. In experiments this transition is observed, the new state being called *modulated wavy vortices*. See Rand [1982], Gorman, Swinney, and Rand [1981], Shaw et al. [1982].

The analysis of the simplest (nondegenerate) $O(2)\times S^1$-symmetric bifurcation problems on the six-dimensional kernel leads to a picture that includes branches corresponding to a variety of the observed flows: Couette, vortices, wavy vortices, twisted vortices, spiral cells, modulated wavy vortices, wavy spirals and an unstable flow found numerically by DiPrima and Grannick [1971] which they call the "non-axisymmetric simple mode." By "correspond" we mean that the solutions we find on the six-dimensional kernel appear to have the same symmetries as the experimentally determined states. As we indicated in the introduction, it is further possible to choose parameters in the model to mimic the observed transition sequences when the outer cylinder is held fixed, and also in the counterrotating case.

DiPrima, Eagles, and Sijbrand [1984] are currently making numerical calculations of certain of the Taylor coefficients of the vector field g obtained by a center manifold reduction. These or similar numerical results should make it possible to determine to what extent the six-dimensional model reflects the expected transitions in the Taylor problem, at least in the counterrotating case.

2. Symmetries on the six-dimensional kernel. Symmetries are introduced in the Taylor problem in three distinct ways:

(1) by the experimental apparatus,

(2) by the mathematical idealization,

(3) by the mathematical analysis.

Since each of these ways introduces a circle group of symmetries the result may seem confusing at first. However, these symmetries do affect the mathematically determined

solutions and, moreover, seem to be present in the experimentally determined states.

The symmetries arising through the apparatus would appear to be the most natural. All formulations of the Taylor problem are invariant under rotation in the azimuthal plane, a plane perpendicular to the cylindrical axis. Rotation through θ in this plane moves one fluid state to another. We denote these symmetries by $SO(2)$.

Next we discuss the symmetries introduced by the mathematical idealization. In the experiments, when vortex flow is observed these vortices tend to have square cross-sections; that is, the height of each vortex is approximately equal to the distance between the cylinders. As a result, in an apparatus whose cylinder length is long compared with the distance between the cylinders, many vortices form at the initial bifurcation. Moreover, the vortex flow appears to be invariant under translation along the cylindrical axis by two band-widths, at least away from the ends of the cylinder. (In the cross-sectional regions vortex flow alternates between clockwise and counterclockwise.)

Thus in the mathematical idealization we assume that the cylinders have infinite length and look only for periodic solutions of period equal to two band-widths. As a result, the Navier–Stokes equations are invariant under both translations along the cylindrical axis and reflection of the cylinder through the azimuthal plane. Periodicity implies that translation by two band-widths acts as the identity. Thus the effective action of this group is by the (compact) group $O(2)$.

Finally, we consider a circle group of symmetries which is introduced into this problem by the technique we use to analyze the bifurcation structure. We use a Lyapunov–Schmidt reduction to determine time-periodic solutions of the Navier–Stokes equations which lie near Couette flow and the parameter values yielding the six-dimensional kernel. The circle group S^1, acting by change of phase on periodic functions, introduces symmetries into this problem. The addition of these S^1 symmetries by the Lyapunov–Schmidt procedure, to the symmetries mentioned above, is described in Sattinger [1983] and Golubitsky and Stewart [1985].

We summarize our discussion here as follows. The full group of symmetries of the Taylor problem on the six-dimensional kernel is:

$$O(2)\times SO(2)\times S^1 \tag{2.1}$$

where $O(2)$ acts by translation and flipping along the cylindrical axis, $SO(2)$ acts by rotation of the azimuthal plane and S^1 acts by change of phase of periodic solutions. For simplicity of notation we assume that the period of the cylindrical translations is 2π and that the period of patterns around the cylinder (in the azimuthal plane) is also 2π. In particular, rotation of the cylinder by half a period is $\pi \in SO(2)$. Moreover, we assume that solutions are 2π-periodic in time.

These assumptions do not affect the group-theoretic formulation of the problem, or its analysis; but they must be correctly interpreted in connection with the observed flows. Since the situation is potentially confusing, a few clarifying remarks may be in order. There is no problem in arranging period 2π for translations: we merely scale the distance along the axis. For periodic solutions in the azimuthal direction a little more caution is required. For example, it is commonly observed in experiments that wavy vortex solutions may appear with wave numbers 3 or 4 (say); that is, with 3 or 4 complete periods relative to a single turn of the cylinder. Provided only *one* such mode is present, we may scale the azimuthal angle to "factor out" this additional periodicity. The angle 2π then represents one period ($2\pi/3$ or $2\pi/4$ on the physical cylinder). In group-theoretic terms, an action of $SO(2)$ for which $\theta \in SO(2)$ produces a rotation by

$k\theta$, k an integer, can be viewed as the standard ($k=1$) action of $SO(2)/Z_k$ and this group may be *identified* with $SO(2)$.

On the six-dimensional kernel, only one such periodic mode occurs, and this procedure may be followed. If two modes with different wave numbers occur, it would be necessary to make $SO(2)$ act by $k\theta$ and $l\theta$ (where k, l are the respective wavenumbers) on the corresponding spaces of eigenfunctions and to carry out the analysis for the appropriate action of $O(2)\times SO(2)\times S^1$. See Chossat [1985].

3. Observed solutions and their symmetries. In the experiments a number of prechaotic states are observed. In this section we discuss the symmetries of each of the following states: Couette flow, Taylor vortices, wavy vortices, spiral cells, and twisted vortices.

Both the *Couette* and *vortex* flows are time-independent. As noted above, vortex flow produces bands along which the flow is in the azimuthal plane. See Fig. 3.1(a).

When $\Omega_0 \leqq 0$, vortex flow loses stability, and a time periodic state called *wavy vortices* appears. See Fig. 3.1(b). This periodic flow has the special form of rotating waves. More precisely, the solution $u(t)$ is a rotating wave if $u(t+\theta)=R_\theta u(t)$ where R_θ denotes rotation by angle θ in the azimuthal plane. We shall see in §4 that *all* periodic solutions obtained from the six-dimensional kernel must be rotating waves.

When $\Omega_0=\Omega$, wavy vortex solutions lose stability, and a new quasi-periodic solution with two independent frequencies appears. This new state is called *modulated wavy vortices*. It is interesting to observe, at this point, how the modulated wavy vortex solution might be detected by our proposed method using a Lyapunov–Schmidt reduction. The idea is to compute the Floquet exponents along the wavy vortex branch of solutions and show that certain of these exponents cross the imaginary axis. Then apply the Sacker–Neimark torus bifurcation theorem to conclude the existence of quasiperiodic solutions. We show in §7 that this scenario is possible. A similar remark holds for identifying wavy spiral states when $\Omega_0<0$.

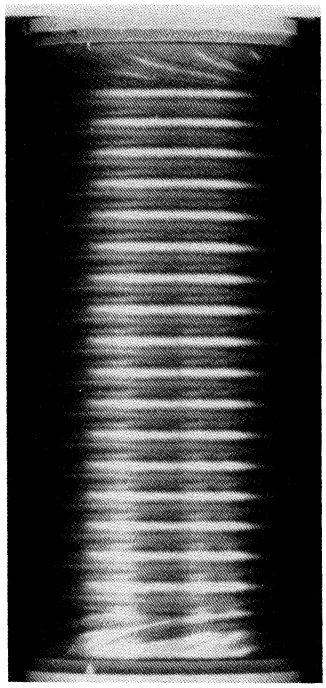
We note that the actual transition to chaos cannot be explained by our analysis. Nevertheless, chaotic behavior may be present in our model and this point deserves further investigation.

We also note here that in the corotating and counterrotating Taylor problems solutions with different planforms are observed. For example, in the corotating case Andereck, Dickman and Swinney [1983] have observed *twisted vortices*. See Fig. 3.1 (c). In the strongly counterrotating case, Couette flow loses stability to a helicoidal pattern called *spiral cells*, which are time-periodic rotating waves. See Fig. 3.1(d).

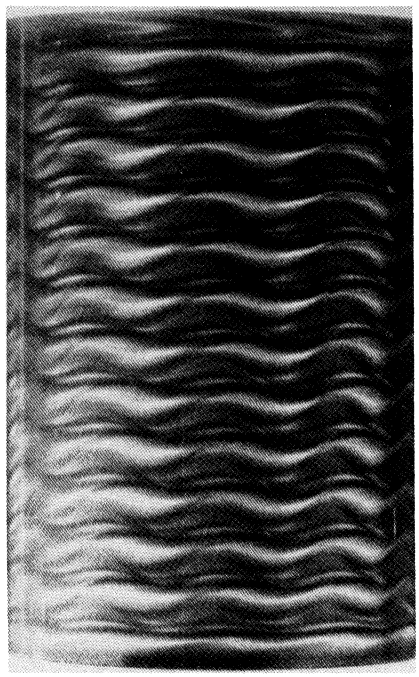
We can distinguish each of the states described above by their *isotropy subgroups*; that is, by the subgroup of (2.1) which leaves the given state invariant. In Table 3.1 we list the isotropy subgroups for each fluid state described above.

We now discuss the entries in Table 3.1. The steady-state solutions are invariant under change of phase (S^1); the periodic solutions are all rotating waves and are invariant under Δ since a change of phase may be compensated for by rotating the cylinder. Couette flow is invariant under all symmetries. Taylor vortices are invariant under all rotations ($SO(2)$) and the flip along the cylindrical axis κ. We denote by $Z_2(\kappa)$ the two-element group generated by κ.

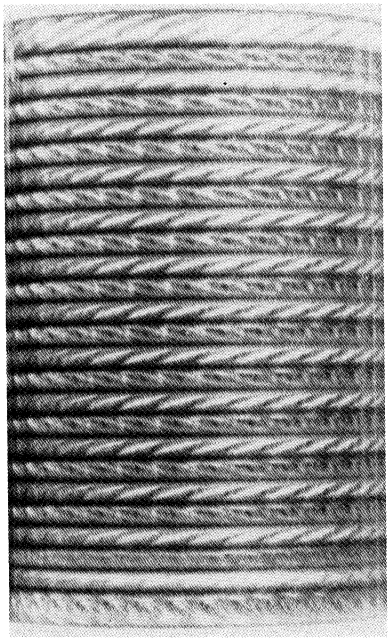
Isotropy subgroups for the periodic solutions are obtained as follows. The helical state, spiral cells, is invariant under $\widetilde{SO(2)}$ since a translation along the cylinder axis may be compensated for by a rotation of the cylinder. Next observe that wavy vortices are invariant under the group element obtained by composing the flip (κ) with rotation



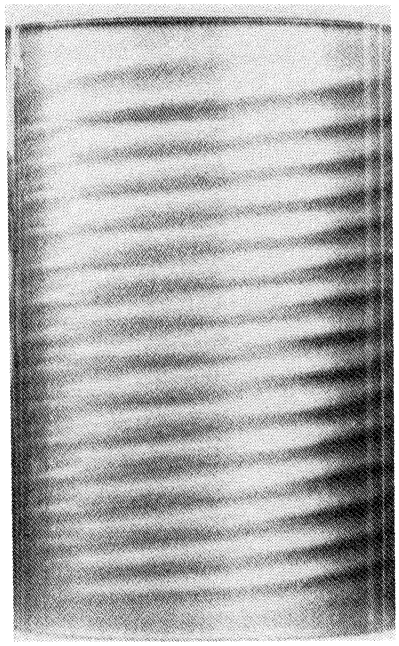
(a) *Taylor vortices* ($R_o = 1164$, $R_i = 1161$).



(b) *Modulated wavy vortices* ($R_o = -100$, $R_i = 350$). *In a still photograph wavy vortices and modulated wavy vortices have a similar appearance.*



(c) *Twisted vortices* ($R_o = 721$, $R_i = 1{,}040$).



(d) *Spiral cells* ($R_o = -295$, $R_i = 237$).

FIG. 3.1. *Observed flows in the Taylor experiment. Reynolds numbers for the inner* (R_i) *and outer* (R_o) *cylinders. Photographs kindly supplied by Harry Swinney and Randy Tagg. Similar photographs will appear in Andereck, Liu and Swinney* [1985].

TABLE 3.1
Isotropy subgroups of observed fluid states.

State	Isotropy subgroup
Couette flow	$O(2)\times SO(2)\times S^1$
Taylor vortices	$Z_2(\kappa)\times SO(2)\times S^1$
Wavy vortices	$Z_2(\kappa,\pi)\times\Delta$
Twisted vortices	$Z_2(\kappa)\times\Delta$
Spiral cells	$\widetilde{SO}(2)\times\Delta$

$\Delta=\{(\theta,-\theta)\in SO(2)\times S^1\}$
$\kappa\in O(2)$ is the flip $z\to -z$ along the cylindrical axis
$\pi\in SO(2)$ is rotation of the azimuthal plane by one-half period
$\widetilde{SO}(2)=\{(\psi,-\psi)\in O(2)\times SO(2)\}$

of the cylinder by half a period $\pi\in SO(2)$. Finally, twisted vortices are invariant under the flip κ.

It is worth noting that the first three bifurcations in the standard Taylor problem ($\Omega_0=0$) break symmetry in a simple way. Couette flow to Taylor vortices breaks the translational symmetries; Taylor vortices to wavy vortices breaks the rotational symmetries ($SO(2)$); and wavy vortices to modulated wavy vortices breaks the rotating wave symmetries (Δ).

4. Group theory and the six-dimensional kernel. In this section, we answer four questions:

(1) What is the exact form of the six-dimensional kernel?

(2) What is the action of the symmetries of the Taylor problem on this kernel?

(3) What is the form of the reduced mapping h, obtained by the Lyapunov–Schmidt procedure?

(4) What are the possible isotropy subgroups of points in the six-dimensional kernel?

We answer the first question by referring to DiPrima and Sijbrand [1982]. Let $\eta=R_i/R_o$ be the ratio of the radii of the inner and outer cylinders. We quote:

> Thus, for example, for $\eta=0.95$ and $\Omega_0/\Omega_i=-0.73976$, Couette flow is simultaneously unstable to an axisymmetric disturbance with wave numbers $(\lambda,m)=(3.482,0)$ and a nonaxisymmetric disturbance with wavenumbers $(\lambda,m)=(3.482,1)$. We also note that...there are 6 critical modes with axial (Z) and azimuthal (Θ) dependence as follows:

$$\cos\lambda Z,\ \sin\lambda Z,\ e^{\pm i\Theta}\cos(\lambda Z),\ e^{\pm i\Theta}\sin(\lambda Z). \tag{4.1}$$

The action of the translations in $O(2)$ on the eigenfunctions in (4.1) is generated by translations of the axial (angle) Z and the flip (κ) which acts by $Z\to -Z$. Rotations in the azimuthal plane act by translations in Θ. (We have omitted the radial dependence of the eigenfunctions here as the group $O(2)\times SO(2)$ acts trivially in the radial direction.) Observe that the resulting action of $O(2)\times SO(2)$ on the six-dimensional space generated by the eigenfunctions in (4.1) leads to the following equivalent action. We identify the six-dimensional kernel with

$$V=\mathbf{R}^2\oplus(\mathbf{R}^2\otimes\mathbf{C}) \tag{4.2}$$

and let elements of $O(2)$ act on $\mathbf{R}^2$ in the standard way and elements of $SO(2)$ act by

multiplication on $\mathbf{C}$. That is

$$(\theta,\psi)(v,w\otimes z)=\big(R_\theta v,(R_\theta w)\otimes(e^{i\psi}z)\big)$$

where R_θ is the usual rotation of $\mathbf{R}^2$ through the angle θ. Similarly, the flip κ acts by $(Kv,Kw\otimes z)$, where K is the matrix $\left(\begin{smallmatrix}1&0\\0&1\end{smallmatrix}\right)$. The action of the phase shifts S^1 in V turns out to be identical with the action of $SO(2)$ on V. This fact can be verified by direct computation. First observe that the $\mathbf{R}^2$ summand in V is spanned by $\{\cos(\lambda Z),\sin(\lambda Z)\}$, a steady-state kernel. As such, S^1 acts trivially on $\mathbf{R}^2$. Next, observe that S^1 commutes with $O(2)\times SO(2)$ and hence the actions of S^1 and $O(2)\times SO(2)$ on $\mathbf{R}^2\otimes\mathbf{C}$ commute. Since the only matrices acting on $\mathbf{R}^2\otimes\mathbf{C}$ commuting with $O(2)\times SO(2)$ are scalar multiples of matrices in $SO(2)$ (see Golubitsky and Stewart [1985, Lemma 3.2]), it follows that the elements of S^1 act in a fashion identical to elements of $SO(2)$. Without loss of generality, we may identify the actions of $SO(2)$ and S^1.

One consequence of this identification is that the subgroup $\Delta=\{(\psi,-\psi)\in SO(2)\times S^1\}$ is in the isotropy subgroup of every element in V. Thus, we have proved:

LEMMA 4.1. *Every periodic solution found in the six-dimensional kernel is a rotating wave.*

A second consequence of the identification of the actions of $SO(2)$ and S^1 is the simple form that the reduction bifurcation equation $h\colon V\to V$, obtained via a Lyapunov–Schmidt reduction, must take. (The function h depends on a number of extra parameters, Ω_0 for example. We suppress this dependence here.) Let the purely imaginary eigenvalues of the linearized Navier–Stokes equations be $\pm\omega i$. For simplicity, use a scaling argument to assume $\omega=1$. Then the idea behind the Lyapunov–Schmidt reduction is to look for small amplitude periodic solutions of period near 2π. One does this by rescaling time in the original equation by a perturbed period parameter τ and looking for precisely 2π-periodic solutions to the scaled equations. What results, after appropriate applications of the implicit function theorem, is a reduction equation

$$h(v,\tau)=0$$

where $h\colon V\times\mathbf{R}\to V$ is smooth and commutes with $O(2)\times SO(2)\times S^1$. We claim that we may assume that the dependence of h on τ is particularly simple. In fact,

$$h(v,\tau)=g(v)-(1+\tau)J \tag{4.3}$$

where J is the matrix form of the action by $\pi/2\in S^1$ on V.

To verify this claim, suppose for the moment that there exists a smooth center manifold. Let $g(v)$ be the reduction vector field on that center manifold. It was proved in Golubitsky and Stewart [1985] that if the Lyapunov–Schmidt reduction is applied to g, introducing τ, then the resulting function h has exactly the form (4.3). This fact relies on having the spatial symmetries $SO(2)$ identified with the temporal symmetries S^1.

If we perform the Lyapunov–Schmidt reduction directly from the Navier–Stokes equations, then the reduced function has the same form as (4.3), at least to first order in τ. In any case, the form (4.3) is used later only to solve certain equations for τ explicitly. If higher order terms are present, then these equations may be solved implicitly, which is sufficient for our purposes. Therefore we lose nothing by working with h in the form (4.3). Moreover, we note that g commutes with the action of $O(2)\times SO(2)\times S^1$ on V and may be identified with the mapping on V obtained by a center manifold reduction, at least up to any finite order in its Taylor expansion.

For the remainder of this section we describe precisely the form that mappings g which commute with $O(2)\times SO(2)\times S^1$ must have. Note that a third consequence of identifying the actions of $SO(2)$ and S^1 is that at this stage we may ignore one of them. Henceforth, we assume that

$$\Gamma = O(2)\times S^1 \tag{4.4}$$

is our group of symmetries and turn attention to the action of Γ on V.

At this point we choose coordinates on V. First write $V = V_1 \oplus V_2$ where $V_1 = \mathbf{R}^2 \cong \mathbf{C}$ and $V_2 = \mathbf{R}^2 \otimes \mathbf{C} \cong M(2,\mathbf{R})$, the space of 2×2 matrices with real entries. Thus elements of V have the form

$$(z,A) \tag{4.5}$$

where

$$z = x + iy \in \mathbf{C} \quad \text{and} \quad A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in M(2,\mathbf{R}).$$

In these coordinates the group action of $\Gamma = O(2)\times S^1$ on (z,A) is defined as follows:

$$(\theta,\psi)(z,A) = \left(e^{i\theta}z, R_\theta A R_\psi\right) \tag{4.6}$$

where

$$R_\theta = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

is the rotation matrix. See Golubitsky and Stewart [1985, §3] for more detail.

We now answer the third question by describing in detail the invariant functions and the equivariant mappings corresponding to this group action. (Recall that Φ: $V \to V$ is equivariant if $\Phi(\gamma v) = \gamma\Phi(v)$ for all $\gamma \in \Gamma$, $v \in V$.) Proofs are found in the Appendix.

PROPOSITION 4.2. *Let ϕ: $V \to R$ be a smooth function defined in a neighborhood of the origin which is invariant with respect to the action of Γ in* (4.6). *Then there exists a smooth function h: $\mathbf{R}^5 \to \mathbf{R}$ defined near* 0 *such that*

$$\phi(v) = h(\beta, N, \delta^2, \gamma, \sigma)$$

where

$$\begin{aligned} &\beta \equiv z\bar{z} = x^2 + y^2, \\ &N = a^2 + b^2 + c^2 + d^2, \\ &\delta = \det A, \\ &\gamma = \operatorname{Re}(z^2\bar{\zeta}), \\ &\sigma = i\delta \operatorname{Im}(z^2\bar{\zeta}) \end{aligned} \tag{4.7}$$

and

$$\zeta = a^2 + b^2 - c^2 - d^2 + 2i(ac + bd).$$

THEOREM 4.3. *Let Φ: $V \to V$ be a smooth Γ-equivariant mapping defined near* 0. *Then there exist Γ-invariant functions*

$$p, q, r, s, P^1, P^2, Q^1, Q^2, Q^3, Q^4, R^1, R^2, R^3, R^4, M^3, M^4$$

such that

$$\Phi(z,A)=\left(pz+qi\delta z+r\bar{z}\zeta+si\delta\bar{z}\zeta,\ \sum_{j=1}^{4}\left(S^jK_j+T^jL_j\right)\right), \tag{4.8}$$

where

$$K_1=\begin{pmatrix}a & b\\ c & d\end{pmatrix},\quad K_2=\begin{pmatrix}-b & a\\ -d & c\end{pmatrix},\quad K_3=\begin{pmatrix}a & b\\ -c & -d\end{pmatrix},\quad K_4=\begin{pmatrix}-b & a\\ d & -c\end{pmatrix},$$
$$L_1=\begin{pmatrix}-c & -d\\ a & b\end{pmatrix},\quad L_2=\begin{pmatrix}d & -c\\ -b & a\end{pmatrix},\quad L_3=\begin{pmatrix}c & d\\ a & b\end{pmatrix},\quad L_4=\begin{pmatrix}-d & c\\ -b & a\end{pmatrix}, \tag{4.9}$$

and

$$\begin{aligned}
&S^1=P^1,\\
&S^2=P^2,\\
&T^1=R^1\delta+Q^1\,\mathrm{Im}(z^2\bar{\zeta}),\\
&T^2=R^2\delta+Q^2\,\mathrm{Im}(z^2\bar{\zeta}),\\
&S^3=Q^3\,\mathrm{Re}(\bar{z}^2)+R^3\delta\,\mathrm{Im}(\bar{\zeta})+M^3\delta\,\mathrm{Im}(\bar{z}^2),\\
&S^4=Q^4\,\mathrm{Re}(\bar{z}^2)+R^4\delta\,\mathrm{Im}(\bar{\zeta})+M^4\delta\,\mathrm{Im}(\bar{z}^2),\\
&T^3=-Q^3\,\mathrm{Im}(\bar{z}^2)+R^3\delta\,\mathrm{Re}(\bar{\zeta})+M^3\delta\,\mathrm{Re}(\bar{z}^2),\\
&T^4=-Q^4\,\mathrm{Im}(\bar{z}^2)+R^4\delta\,\mathrm{Re}(\bar{\zeta})+M^4\delta\,\mathrm{Re}(\bar{z}^2).
\end{aligned} \tag{4.10}$$

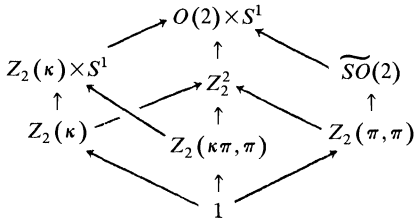
We shall exploit the form of ϕ in (4.8) to solve explicitly the reduced bifurcation equation $g=0$. In order to understand what types of solutions one may find in $g=0$ we answer the fourth question of this section. By determining, up to conjugacy, the set of all isotropy subgroups of elements in V, we determine the symmetries that possible solutions to the Navier–Stokes equations found by reducing to V may have.

The lattice (of conjugacy classes) of isotropy subgroups for Γ acting on V is given in Table 4.1. Containment of one conjugacy class in another is indicated by arrows. In Table 4.2 we list these isotropy subgroups along with the states in the Taylor problem which have those symmetries. We use the notation Z_2 to indicate a two-element group and $Z_2(\alpha)$ to indicate the two-element group generated by $\alpha\in\Gamma$.

We emphasize that the containments in Table 4.1 are of conjugacy classes. For example, $Z(\kappa\pi,\pi)$ is not contained in $Z_2(\kappa)\times S^1$. However, $Z_2(\kappa\pi,\pi)$ is conjugate to $Z_2(\kappa,\pi)$ which is contained in $Z_2(\kappa)\times S^1$.

TABLE 4.1

Lattice of conjugacy classes of isotropy subgroups of Γ *acting on* V.



Note: Z_2^2 is generated by κ, $(\kappa\pi,\pi)$, (π,π).

TABLE 4.2

The symmetries associated with observed fluid states.

Isotropy subgroup	Solution type
$O(2)\times S^1$	Couette flow
$Z_2(\kappa)\times S^1$	Taylor vortices
$\widetilde{SO}(2)$	spiral cells
$Z_2(\kappa\pi,\pi)$	wavy vortices
$Z_2(\kappa)$	twisted vortices

We derive the lattice pictured in Table 4.1 by first considering orbit representatives. Begin by considering the action of $O(2)\times S^1$ on $\mathbf{R}^2\otimes\mathbf{C}\cong M(2,\mathbf{R})$. Let A be a 2×2 matrix. As shown in Golubitsky and Stewart [1985, §7], we can choose an element of $O(2)\times S^1$ so that A is conjugated to the diagonal matrix $\left(\begin{smallmatrix} a & 0 \\ 0 & d \end{smallmatrix}\right)$ where $a\geqq d\geqq 0$. It is then easy to show that there are four types of orbits as shown in Table 4.3.

TABLE 4.3

Orbit representatives of $O(2)\times S^1$ acting on $M(2,\mathbf{R})$.

Orbit representative	Isotropy subgroup
0	$O(2)\times S^1$
$\left(\begin{smallmatrix} a & 0 \\ 0 & 0 \end{smallmatrix}\right)$, $a>0$	Z_2^2
$\left(\begin{smallmatrix} a & 0 \\ 0 & a \end{smallmatrix}\right)$, $a>0$	$\widetilde{SO}(2)$
$\left(\begin{smallmatrix} a & 0 \\ 0 & d \end{smallmatrix}\right)$, $a>d>0$	$Z_2(\pi,\pi)$

Having put the matrices in $M(2,\mathbf{R})$ into normal form, we now use the isotropy subgroups of these matrices to conjugate the elements $z\in\mathbf{C}\cong\mathbf{R}^2$. In this way we obtain representatives for all the orbits of $O(2)\times S^1$ acting on $\mathbf{R}^2\oplus(\mathbf{R}^2\otimes\mathbf{C})$. These results are summarized in Table 4.4.

TABLE 4.4

Orbit representatives of $O(2)\times S^1$ acting on $\mathbf{R}^2\oplus(\mathbf{R}^2\otimes\mathbf{C})$.

Orbit representative (z,A)	Isotropy subgroup
$(0,0)$	$O(2)\times S^1$
$(x,0)$, $x>0$	$Z_2(\kappa)\times S^1$
$\left(0,\left(\begin{smallmatrix} a & 0 \\ 0 & 0 \end{smallmatrix}\right)\right)$, $a>0$	Z_2^2
$\left(x,\left(\begin{smallmatrix} a & 0 \\ 0 & 0 \end{smallmatrix}\right)\right)$, $x>0$, $a>0$	$Z_2^2(\kappa)$
$\left(iy,\left(\begin{smallmatrix} a & 0 \\ 0 & 0 \end{smallmatrix}\right)\right)$, $y>0$, $a>0$	$Z_2^2(\kappa\pi,\pi)$
$\left(x+iy,\left(\begin{smallmatrix} a & 0 \\ 0 & 0 \end{smallmatrix}\right)\right)$, $x>0$, $y>0$, $a>0$	1
$\left(0,\left(\begin{smallmatrix} a & 0 \\ 0 & a \end{smallmatrix}\right)\right)$, $a>0$	$\widetilde{SO}(2)$
$\left(x,\left(\begin{smallmatrix} a & 0 \\ 0 & a \end{smallmatrix}\right)\right)$, $x>0$, $a>0$	1
$\left(0,\left(\begin{smallmatrix} a & 0 \\ 0 & d \end{smallmatrix}\right)\right)$, $a>d>0$	$Z_2(\pi,\pi)$
$\left(x+iy,\left(\begin{smallmatrix} a & 0 \\ 0 & d \end{smallmatrix}\right)\right)$, $\lvert x+iy\rvert\neq 0$, $x\geqq 0$, $a>d>0$	1

TABLE 4.5

Fixed-point subspaces of $O(2)\times S^1$ acting on V.

Isotropy subgroup	Fixed-point subspace	Dimension
$O(2)\times S^1$	0	0
$Z_2(\kappa)\times S^1$	$(x,0)$	1
Z_2^2	$\left(0,\begin{pmatrix} a & b \\ 0 & 0 \end{pmatrix}\right)$	2
$\widetilde{SO}(2)$	$\left(0,\begin{pmatrix} a & b \\ -b & a \end{pmatrix}\right)$	2
$Z_2(\kappa)$	$\left(x,\begin{pmatrix} a & b \\ 0 & 0 \end{pmatrix}\right)$	3
$Z_2(\kappa\pi,\pi)$	$\left(iy,\begin{pmatrix} a & b \\ 0 & 0 \end{pmatrix}\right)$	3
$Z_2(\pi,\pi)$	$\left(0,\begin{pmatrix} a & b \\ c & d \end{pmatrix}\right)$	4
1	$\left(x+iy,\begin{pmatrix} a & b \\ c & d \end{pmatrix}\right)$	6

In the last table of this section we present the fixed-point subspaces of the various isotropy subgroups. More precisely, let $\Sigma\subset\Gamma$ be a subgroup. Define

$$V^{\Sigma}=\{v\in V \mid \sigma v=v \text{ for all } \sigma\in\Sigma\}. \tag{4.11}$$

Observe that if Φ: $V\to V$ commutes with Γ, then Φ maps V^{Σ} to itself (see Golubitsky and Stewart [1985, (1.6)]).

5. Branching and stability. Let $h(z,A,\lambda,\tau)$ be the mapping on the six-dimensional kernel obtained via the Lyapunov–Schmidt reduction. Note that h depends explicitly on the bifurcation parameter λ and the perturbed period τ. In addition, we know that h commutes with the symmetries in the Taylor problem. We shall use the consequences of this fact to explain how to compute both the solutions to $h=0$ and the eigenvalues of the 6×6 Jacobian matrix dh along branches of solutions to $h=0$. One consequence of the $O(2)\times SO(2)\times S^1$ symmetries is that the eigenvalues of dh determine the orbital asymptotic stability of solutions.

Let us be more precise. Recall the form of h in (4.3) with its simple τ-dependence, namely

$$h(z,A,\lambda,\tau)=g(z,A,\lambda)-(1+\tau)\left(0,\begin{pmatrix} -b & a \\ -d & c \end{pmatrix}\right) \tag{5.1}$$

where $A=\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$. In deriving this form we note that if the Navier–Stokes equations admit a smooth center manifold then h will have exactly the form (5.1), where g is the reduced vector field on that center manifold. We proved in Golubitsky and Stewart [1985, Thm. 8.2] that the eigenvalues of dh determine the orbital asymptotic stability of solutions to the vector field g on the center manifold. Moreover the center manifold reduction implies that the stabilities of solutions to the vector field g are the same as the stabilities of the corresponding solutions to the Navier–Stokes equations.

If there should not exist a smooth center manifold (which we doubt) then we are computing the correct stabilities for g, accurate to any finite order, by using the eigenvalues of $(dh)\,|_{h=0}$.

We now describe how to compute the eigenvalues of dh. Recall that (4.8) restricts the form of g in (5.1) to:

$$g(z,A,\lambda)=\left(pz+qi\delta z+r\bar{z}\zeta+si\delta\bar{z}\zeta,\ \sum_{j=1}^{4}S^jK_j+T^jL_j\right) \tag{5.2}$$

where p,q,r,s and the coefficients $P^1,\cdots,M^4$ appearing in (4.11) are invariant functions, hence functions of the five variables $\beta,N,\delta^2,\gamma,\sigma$ defined in (4.7) and λ. Moreover, since h is obtained via the Lyapunov–Schmidt reduction, the linear terms must vanish. Hence

$$p(0)=0,\quad P^1(0)=0,\quad P^2(0)=1. \tag{5.3}$$

Equivariance shows that to solve the equations $h=0$ we need only evaluate h on typical orbit representatives. The resulting equations are listed in Table 5.1. See Table 4.4 for the list of orbit representatives and their isotropy subgroups.

Remarks. (i) The equations involving τ serve only to determine the perturbed period of the associated periodic solution. Note that (5.1) allows us to eliminate τ by solving these equations explicitly (or implicitly if there does not exist a smooth center manifold, see §4 above).

(ii) Observe that τ is indeterminate on the $Z_2(\kappa)\times S^1$ branch, which is to be expected since the bifurcation is to a "steady state."

(iii) Observe that the theoretical basis of our explicit calculations is given by (4.10): fixed-point subspaces V^Σ are mapped to themselves by equivariant mappings. Therefore we may restrict h to V^Σ and seek solutions to $h\,|\,V^\Sigma=0$, considering each isotropy subgroup Σ in turn.

Table 5.1 also lists the coefficients that determine the signs of the (real parts of the) eigenvalues of dh. We consider the branching equations briefly first, and then describe in more detail the eigenvalue calculations.

By writing (4.9) in coordinates we obtain

$$h\left(x,y,\begin{pmatrix}a & b\\ c & d\end{pmatrix}\right)=\left(X,Y,\begin{pmatrix}A & B\\ C & D\end{pmatrix}\right), \tag{5.4}$$

where

$$\begin{aligned}
&\text{(a)}\ X=px-q\delta y+rx\,\mathrm{Re}(\zeta)+ry\,\mathrm{Im}(\zeta)+s\delta y\,\mathrm{Re}(\zeta)-s\delta x\,\mathrm{Im}(\zeta),\\
&\text{(b)}\ Y=py+q\delta x-ry\,\mathrm{Re}(\zeta)-rx\,\mathrm{Im}(\zeta)+s\delta x\,\mathrm{Re}(\zeta)+s\delta y\,\mathrm{Im}(\zeta),\\
&\text{(c)}\ A=(S^1+S^3)a+(-S^2-S^4)b+(-T^1+T^3)c+(T^2-T^4)d,\\
&\text{(d)}\ B=(S^2+S^4)a+(S^1+S^3)b+(-T^2+T^4)c+(-T^1+T^3)d,\\
&\text{(e)}\ C=(T^1+T^3)a+(-T^2-T^4)b+(S^1-S^3)c+(-S^2+S^4)d,\\
&\text{(f)}\ D=(T^2+T^4)a+(T^1+T^3)b+(S^2-S^4)c+(S^1-S^3)d.
\end{aligned} \tag{5.5}$$

The branching equations always take the form $X=Y=A=B=C=D=0$, evaluated on the appropriate orbit representative. The entries in the table follow readily. However, a few comments should be made regarding the last four entries of "unknown" type.

TABLE 5.1

Branching equations and eigenvalues for solutions with given symmetry.

Solution type; Isotropy; orbit	Branching equations (to be evaluated at orbit representative shown)	Signs of eigenvalues	Multiplicity
Couette flow $O(2)\times S^1$ $(0,0)$	None	p $P^1 \pm i(P^2-1-\tau)$	2 2, 2
Taylor vortices $Z_2(\kappa)\times S^1$ $(x,0)$	$p=0$ at: $(x^2,0,0,0,0,\lambda)$[1]	0 p_β $(P^1+x^2Q^3)\pm i(P^2-1-\tau+x^2Q^4)$ $(P^2-x^2Q^3)\pm i(P^2-1-\tau-x^2Q^4)$	1 1 1, 1 1, 1
unknown Z_2^2 $\left(0,\begin{pmatrix} a & 0 \\ 0 & 0 \end{pmatrix}\right)$	$P^1=0$ $P^2=1+\tau$ at: $(0,a^2,0,0,0,\lambda)$	0 $p+ra^2$ $P_N^1+a^2P_N^3$ $R^2+a^2R^4$	2 1 1 1 1
spiral cells $\widetilde{SO}(2)$ $\left(0,\begin{pmatrix} a & 0 \\ 0 & a \end{pmatrix}\right)$	$P^1+a^2R^2=0$ $P^2-a^2R^1=1+\tau$ at: $(0,2a^2,a^4,0,0,\lambda)$	0 $p\pm iqa^2$ $2P_N^1+R^2+a^2(P_{\delta^2}^1+2R_N^2)+a^4R_{\delta^2}^2$ $(R^2+2a^2R^4)\pm i(R^1+2a^2R^4)$	1 1, 1 1 1, 1
wavy vortices $Z_2(\kappa\pi,\pi)$ $\left(iy,\begin{pmatrix} a & 0 \\ 0 & 0 \end{pmatrix}\right)$	$p-a^2r=0$ $P^1-y^2Q^3=0$ $P^2-y^2Q^4=1+\tau$ at: $(y^2,a^2,0,-a^2y^2,0,\lambda)$	0 r $R^2a^2+2Q^3y^2+ay^2Q_d^4$ $+a^4R^4-a^2y^2M^4$ $\begin{bmatrix} Y_y & Y_a \\ A_y & A_a \end{bmatrix}^{[2]}$ $\det = p_\beta P_N^1-(p_N-r)(P_\beta^1-Q^3)+\cdots$ $\text{trace} = p_\beta y^2+P_N^1a^2+\cdots$	2 1 1 1, 1
twisted vortices $Z_2(\kappa)$ $\left(x,\begin{pmatrix} a & 0 \\ 0 & 0 \end{pmatrix}\right)$	$p+a^2r=0$ $P^1+x^2Q^3=0$ $P^2+x^2Q^4=1+\tau$ at: $(x^2,a^2,0,a^2x^2,0,\lambda)$	0 $-r$ $R^2a^2-2Q^3x^2$ $-ax^2Q_d^4+a^4R^4+a^2x^2M^4$ $\begin{bmatrix} X_x & X_a \\ A_x & A_a \end{bmatrix}^{[2]}$ $\det = p_\beta P_N^1-(p_N+r)(P_\beta^1-Q^3)+\cdots$ $\text{trace} = p_\beta x^2+P_N^1a^2+\cdots$	2 1 1 1, 1
unknown $Z_2(\pi,\pi)$ $\left(0,\begin{pmatrix} a & 0 \\ 0 & d \end{pmatrix}\right)$	$R^2+(a^2+d^2)R^4=0$ plus others at: $(0,a^2+d^2,a^2d^2,0,0,\lambda)$	not computed	

TABLE 5.1 (continued)

unknown 1 $\left(x,\begin{pmatrix} a & 0 \\ 0 & a \end{pmatrix}\right)$	$q=0$ plus others at: $(x^2, 2a^2, a^4, 0, 0, \lambda)$	not computed	
unknown 1 $\left(x+iy,\begin{pmatrix} a & 0 \\ 0 & 0 \end{pmatrix}\right)$	$r=0$ plus others at: $(x^2+y^2, a^2, 0, (x^2-y^2)a^2, 0, \lambda)$	not computed	
unknown 1 $\left(x+iy,\begin{pmatrix} a & 0 \\ 0 & d \end{pmatrix}\right)$	$X=Y=A=B=C=D=0$ at: $(x^2+y^2, a^2+d^2, a^2d^2,$ $(x^2-y^2)(a^2-d^2), 2xyad(a^2-d^2), \lambda)$	not computed	

Notes: [1] The equations must be evaluated at $(\beta, N, \delta^2, \gamma, \sigma, \lambda)$ where these in turn are evaluated on an orbit representative to yield the form stated.

[2] The remaining eigenvalues are those of the specified 2×2 matrix. Its determinant and trace are shown to lowest order (omitting positive factors) to determine their signs.

When the isotropy group is $Z_2(\pi,\pi)$ we take two of the equations, namely $A=0=D$, which reduce to:

$$(S^1+S^3)a+(T^2-T^4)d=0,$$

$$(T^2+T^4)a+(S^1-S^3)d=0.$$

Now observe that S^1+S^3 and T^2+T^4 have a factor d. Divide this out and subtract. The result has a factor (a^2-d^2), and the entry in the table follows. We do not require the remaining equations, because an appeal to nondegeneracy (§6) now rules out this case.

For the next two cases, the equations $X=Y=0$ lead, among other things, to the listed equation, which is also ruled out by nondegeneracy. The final case is extremely complicated and it remains possible that such a branch might occur: see §7 for further discussion.

The computation of the eigenvalues, particularly those along the $Z_2(\kappa)$ and $Z_2(\kappa\pi,\pi)$ branches, is the most difficult part of this section. This computation is facilitated by the use of several results in Golubitsky and Stewart [1985, §8b]. The first is that along a solution branch (v_0,λ_0,τ_0) with isotropy subgroup Σ, the derivative $(dh)_{v_0,\lambda_0,\tau_0}$ commutes with Σ. This implies (Lemma 8.4 of that paper) that $(dh)_{v_0,\lambda_0,\tau_0}$ leaves invariant the subspaces W_j of V formed by adding together all subspaces of V on which Σ acts by a fixed irreducible representation. We use the W_j to put dh into block diagonal form.

The second fact is that $(dh)_{v,\lambda,\tau}$ vanishes on all vectors tangent to the orbit of v under the action of $O(2)\times S^1$. These null-vectors are given by:

$$\text{(5.6)}\qquad \text{(a)}\quad (z,A)\to\frac{d}{d\theta}(z,AR_\theta)\Big|_{\theta=0}=\left(0,A\begin{pmatrix}0 & -1\\ 1 & 0\end{pmatrix}\right),$$

$$\text{(b)}\quad (z,A)\to\frac{d}{d\psi}\left(e^{i\psi}z,R_\psi A\right)\Big|_{\psi=0}=\left(iz,\begin{pmatrix}0 & -1\\ 1 & 0\end{pmatrix}A\right).$$

We now outline the explicit computation of the eigenvalues listed in Table 5.1.

(a) $Z_2 \times S^1$ (*Taylor vortices*). Decompose $V = \mathbf{R}^2 \oplus (\mathbf{R}^2 \otimes \mathbf{C})$ into irreducibles for $Z_2 \times S^1$. We get $V = W_0 \oplus W_1 \oplus W_2 \oplus W_3$ where

$$W_0 = \{(x,0)\}, \qquad W_2 = \left\{\left(0, \begin{pmatrix} a & b \\ 0 & 0 \end{pmatrix}\right)\right\},$$

$$W_1 = \{(iy,0)\}, \qquad W_3 = \left\{\left(0, \begin{pmatrix} 0 & 0 \\ c & d \end{pmatrix}\right)\right\}.$$

The actions are given by

	W_0	W_1	W_2	W_3
$\kappa \in O(2)$	1	-1	1	-1
$\psi \in SO(2)$	1	1	$\mathbf{R}_\psi$	$\mathbf{R}_\psi$

which are distinct irreducible actions. Therefore, dh leaves each W_j invariant. Let $\Phi_j = dh \mid W_j$, so that dh has the block form:

	x	y	a b	c d
x	Φ_0	0	0	0
y	0	Φ_1	0	0
a b	0	0	Φ_2	0
c d	0	0	0	Φ_3

We evaluate the Φ_j at the orbit representative $(x,0)$, with the following results.

$\Phi_0 = X_x = p + p_x x = p_x x$ since $p = 0$ on this branch by Table 5.1. Now $x > 0$ so we can divide it out.

$$\Phi_1 = Y_y = p + p_y y = 0,$$

$$\Phi_2 = \begin{bmatrix} A_a & A_b \\ B_a & B_b \end{bmatrix} = \begin{bmatrix} S^1 + S^3 & -S^2 - S^4 \\ S^2 + S^4 & S^1 + S^3 \end{bmatrix} = \begin{bmatrix} P^1 + x^2 Q^3 & -(P^2 - 1 - \tau + x^2 Q^4) \\ P^2 - 1 - \tau + x^2 Q^4 & P^1 + x^2 Q^3 \end{bmatrix}.$$

Since a matrix $\left(\begin{smallmatrix} \alpha & -\beta \\ \beta & \alpha \end{smallmatrix}\right)$ has eigenvalues $a \pm i\beta$, the entry in the table follows. Similarly

$$\Phi_3 = \begin{bmatrix} C_c & C_d \\ D_c & D_d \end{bmatrix} = \begin{bmatrix} S^1 - S^3 & -S^2 + S^4 \\ S^2 - S^4 & S^1 - S^3 \end{bmatrix} = \begin{bmatrix} P^1 - x^2 Q^3 & -(P^2 - 1 - \tau - x^2 Q^4) \\ P^2 - 1 - \tau - x^2 Q^4 & P^1 - x^2 Q^3 \end{bmatrix}.$$

The entries for vortices in Table 5.1 follow. Note that Φ_2 and Φ_3 have to be scalar multiples $\left(\begin{smallmatrix} \alpha & -\beta \\ \beta & \alpha \end{smallmatrix}\right)$ of rotations since dh commutes with S^1; direct calculation confirms this.

(b) Z_2^2 (*unknown*). We use the same decomposition $V = W_0 \oplus W_1 \oplus W_2 \oplus W_3$ as in (a). The actions are:

	W_0	W_1	W_2	W_3
κ	1	-1	1	-1
(π,π)	-1	-1	1	1

So the irreducible components on these spaces are distinct. Therefore dh leaves each W_j invariant: set $\Phi_j = dh \mid W_j$. We have

$$\Phi_0 = p + ra^2, \qquad \Phi_1 = p - ra^2.$$

Now dh has one eigenvalue 0 on each of W_2 and W_3, so the remaining eigenvalues are $\operatorname{Tr}\Phi_2$, $\operatorname{Tr}\Phi_3$. Use (5.4) here. These are computed as follows.

$$\operatorname{Tr}\Phi_2 = A_a + B_b = (S^1 + S^3) + (S_a^1 + S_a^3)a + (S^1 + S^3) + (S_b^2 + S_b^4)a$$

since $b = c = d = 0$ on the orbit. The branching equations show that $S^1 + S^3 = 0$, so the sign is given by $S_a^1 + S_a^3 + S_b^2 + S_b^4$. Now on the orbit we have $z = 0$, $\delta = 0$, $\zeta = a^2$, $\delta_a = 0$, $\delta_b = 0$. So by (4.11) we compute this as $P_a^1 + P_b^2$. But $N_b = 0$ on the orbit, so this becomes $P_N^1 \cdot 2a$. Dividing by $2a > 0$ gives the table entry. (From now on, we omit such details from the calculation.)

Similarly,

$$\operatorname{Tr}\Phi_3 = C_c + D_d = S^1 - S^2 + (T_c^1 + T_c^3)a + (S^1 - S^3) + (T_d^2 + T_d^4)a.$$

But $S^1 + S^3 = 0$ on this branch, so this is

$$4S^1 + a(T_c^1 + T_c^3 + T_d^2 + T_d^4) = a^2R^2 + a^4R^4.$$

(c) $\widetilde{SO}(2)$ (*spiral cells*). Let

$$W_0 = \mathbf{C}, \quad W_1 = \left\{\begin{pmatrix} a & b \\ -b & a \end{pmatrix}\right\}, \quad W_2 = \left\{\begin{pmatrix} a & b \\ b & -a \end{pmatrix}\right\}.$$

These are irreducible under $\widetilde{SO}(2)$; and $(\theta, -\theta) \in \widetilde{SO}(2)$ acts on W_0 as $e^{i\theta}$, on W_1 as the identity, and on W_2 as $e^{2i\theta}$ (see Golubitsky and Stewart [1985, (10.5)]). Hence the W_j are invariant under dh. Let $\Phi_j = dh \mid W_j$ as usual. We compute

$$\Phi_0 = \begin{bmatrix} p & -qa^2 \\ qa^2 & p \end{bmatrix},$$

which has the form $\left(\begin{smallmatrix} \alpha & -\beta \\ \beta & \alpha \end{smallmatrix}\right)$ required to commute with $\widetilde{SO}(2)$. Now Φ_1 has one zero eigenvalue on W_1, so the other is

$$\operatorname{Tr}\Phi_1 = A_a + A_d + B_b - B_c$$

by Golubitsky and Stewart [1985, (10.11)]. We compute this on the orbit $(0, \left(\begin{smallmatrix} a & 0 \\ 0 & a \end{smallmatrix}\right))$. The result is

$$\operatorname{Tr}\Phi_1 = 2a(S_a^1 + S_a^3 + T_a^2 - T_a^4 + S_d^1 + S_d^4 + T_d^2 - T_d^4).$$

In deriving this, note that $S^1 + S^3 + T^3 - T^4 = 0$ by the branching equations. Also, the b- and c-derivatives of the invariants are equal, so the b- and c-derivative terms cancel. The a- and d-derivatives of β, N, δ, γ, σ are equal, whereas $\zeta_a = 2a = -\zeta_d$. Hence the only terms that remain, on dividing out positive factors, are

$$2P_N^1 + R^2 + a^2(P_{\delta^2}^1 + 2R_N^2) + a^4R_{\delta^2}^2.$$

The matrix of Φ_2 can be computed as

$$\begin{bmatrix} A_a - A_d & A_b + A_c \\ B_a - B_d & B_b + B_c \end{bmatrix}$$

and must be of the form $\begin{pmatrix} \alpha & -\beta \\ \beta & \alpha \end{pmatrix}$ by Golubitsky and Stewart [1985, §10]. Much as above, we find that

$$A_a - A_d = -2a^2(R^2 + 2a^2R^4),$$
$$A_b + A_c = -2a^2(R^1 + 2a^2R^4),$$

as required for the entry in the table.

(d) $Z_2(\kappa\pi, \pi)$ (*wavy vortices*). We decompose $V = W_0 \oplus W_1$ where

$$W_0 = \langle y, a, b \rangle = +1 \text{ eigenspace},$$
$$W_1 = \langle x, c, d \rangle = -1 \text{ eigenspace},$$

and take a basis in the order

$$y, a, b; x, c, d.$$

Let $\Phi_j = dh \mid W_j$. The null-vectors for the two zero eigenvalues of dh may be found from (5.4) and are

$$\begin{bmatrix} 0 \\ 0 \\ -a \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 0 \\ 0 \\ 0 \\ -y \\ a \\ 0 \end{bmatrix}$$

with respect to this basis, when evaluated on the orbit. So column b of dh is zero and columns x, c are linearly dependent. Direct calculation yields $C_x = D_x = 0$, whence by linear dependence $C_c = D_c = 0$. So dh is of the form

$$\begin{array}{c} \\ y \\ a \\ b \\ \hline x \\ c \\ d \end{array} \begin{array}{c} \begin{array}{cccccc} y & a & b & x & c & d \end{array} \\ \left[\begin{array}{cccccc} * & * & 0 & 0 & 0 & 0 \\ * & * & 0 & 0 & 0 & 0 \\ * & * & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & * & * & * \\ 0 & 0 & 0 & 0 & 0 & * \\ 0 & 0 & 0 & 0 & 0 & * \end{array}\right] \end{array}.$$

The eigenvalues of Φ_2 are therefore

$$0,$$
$$X_x = 2ra^2,$$
$$D_d = a^2R^2 + 2y^2Q^3 + ay^2Q_d^4 + a^4R^4 - a^2y^2M^4,$$

and those of Φ_1 are given by

$$\begin{bmatrix} Y_y & Y_a \\ A_y & A_a \end{bmatrix}.$$

Now

$$
\begin{aligned}
Y_y &= p + p_y y - ra^2 - r_y y a^2 = p_y y - r_y a^2 y, \\
Y_a &= p_a y - \left(a^2 r_a + 2ar\right) y, \\
A_y &= \left(S_y^1 + S_y^3\right) a, \\
A_a &= \left(S_a^1 + S_a^3\right) a \quad \text{since } S^1 + S^3 = 0.
\end{aligned}
$$

To evaluate the y- and a-derivatives, note that on the orbit,

$$
\begin{aligned}
&N_y = 0, \quad \beta_y = 2y, \quad (\delta^2)_y = 0, \quad \gamma_y = -2ya^2, \quad \sigma_y = 0, \\
&N_a = 2a, \quad \beta_a = 0, \quad (\delta^2)_a = 0, \quad \gamma_a = -2ay^2, \quad \sigma_a = 0.
\end{aligned}
$$

The y-derivatives introduce a factor y, the a-derivatives a factor a. So the matrix is of the form

$$
\begin{bmatrix} e_{11} y^2 & e_{12} ay \\ e_{21} ay & e_{22} a^2 \end{bmatrix}
$$

for certain functions e_{ij}. We therefore evaluate the e_{ij} to lowest order. The result is

$$
\begin{aligned}
\begin{bmatrix} e_{11} y^2 & e_{12} ay \\ e_{21} ay & e_{22} a^2 \end{bmatrix} &= \begin{bmatrix} p_y y & p_a y - 2ary \\ \left(S_y^1 + S_y^3\right) a & \left(S_a^1 + S_a^3\right) a \end{bmatrix} \\
&= \begin{bmatrix} 2p_\beta y^2 & 2\left(p_N - r\right) ay \\ 2\left(P_\beta^1 - Q^3\right) ay & 2P_N^1 a^2 \end{bmatrix}.
\end{aligned}
$$

The determinant and trace of Φ_1 therefore have the signs indicated in the table.

(e) $Z_2(\kappa)$ (*twisted vortices*). This is similar to (d). The decomposition into invariant subspaces is now $V = W_0 \oplus W_1$ where

$$
\begin{aligned}
W_0 &= \langle x, a, b \rangle = +1 \text{ eigenspace for } \kappa, \\
W_1 &= \langle y, c, d \rangle = -1 \text{ eigenspace for } \kappa.
\end{aligned}
$$

We take a basis for V in the order

$$
x, a, b; y, c, d.
$$

Then dh has block form, and we let $dh \mid W_j = \Phi_j$.

There are two zero eigenvalues of dh given by (5.4). In the basis above the associated eigenvectors are

$$
\begin{bmatrix} x \\ a \\ b \\ y \\ c \\ d \end{bmatrix} \to \begin{bmatrix} 0 \\ b \\ -a \\ 0 \\ d \\ -c \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} -y \\ -c \\ -d \\ x \\ a \\ b \end{bmatrix}.
$$

Evaluation on the orbit representative $(x, \left(\begin{smallmatrix} a & 0 \\ 0 & 0 \end{smallmatrix}\right))$ yields the eigenvectors

$$\begin{bmatrix} 0 \\ 0 \\ -a \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 0 \\ 0 \\ 0 \\ x \\ a \\ 0 \end{bmatrix}.$$

Therefore column b of dh is zero and columns y and c are linearly dependent. Putting the zeros in column b, we get

$$d\phi = \begin{array}{c} \\ x \\ a \\ b \\ y \\ c \\ d \end{array} \left[\begin{array}{ccc|ccc} x & a & b & y & c & d \\ * & * & 0 & 0 & 0 & 0 \\ * & * & 0 & 0 & 0 & 0 \\ * & * & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & \multicolumn{3}{c}{\multirow{3}{*}{Φ_1}} \\ 0 & 0 & 0 & & & \\ 0 & 0 & 0 & & & \end{array}\right].$$

Direct calculation, evaluating at $y=b=c=d=0$, yields $C_y=0$, $D_y=0$, whence by linear dependence of columns y and c we also have $C_c=D_c=0$. So

$$\Phi_1 = \left[\begin{array}{cc|c} Y_y & Y_c & Y_d \\ \hline 0 & 0 & C_d \\ 0 & 0 & D_d \end{array}\right].$$

This is triangular, so its eigenvalues are

$$0,$$
$$Y_y = p - a^2 r = -2ra^2,$$
$$D_d = a^2R^2 - 2x^2Q^3 - ax^2Q_d^4 + a^4R^4 + a^2x^2M^4,$$

using the branching equations.

Since column b of Φ_0 is zero, the eigenvalues of Φ_0 are 0, together with those of

$$\begin{bmatrix} X_x & X_a \\ A_x & A_a \end{bmatrix}.$$

Now

$$X_x = p_x x + r_x x a^2,$$
$$X_a = p_a x + \left(a^2 r_a + 2ar\right)x,$$
$$A_x = \left(S_x^1 + S_x^3\right)a,$$
$$A_a = \left(S_a^1 + S_a^3\right)a.$$

Again the matrix is of the form

$$\begin{bmatrix} f_{11}y^2 & f_{12}ay \\ f_{21}ay & f_{22}a^2 \end{bmatrix}$$

and we may retain only the lowest order terms in the f_{ij}. The result is

$$\begin{bmatrix} 2p_\beta x^2 & 2(p_N+r)ax \\ 2(P^1_\beta - Q^3)ax & 2P^1_N a^2 \end{bmatrix}$$

so the determinant and trace have the indicated signs.

Notice the "duality" between wavy and twisted vortices. This completes the verification of Table 5.1.

6. Nondegeneracy conditions. We now proceed to a detailed analysis of the solutions to the branching equations, and the signs of the real parts of the eigenvalues along branches, which determine (orbital asymptotic) stability. The main qualitative features of the bifurcation diagrams, and their associated stabilities, depend upon the signs of a number of coefficients. We therefore impose appropriate *nondegeneracy conditions*: these coefficients should be nonzero.

Recall from §1 that in order to obtain the six-dimensional kernel we had to fix the speed of counterrotation of the outer cylinder at some critical value Ω_0^*. At this speed we found a two-dimensional eigenspace associated with zero eigenvalues, $\mathbf{R}^2$, coalescing with a four-dimensional space associated with a pair of complex conjugate purely imaginary eigenvalues, $\mathbf{R}^2 \otimes \mathbf{C}$. Thus, in order to model the effects of counterrotation in the Taylor experiment, we must introduce a perturbation parameter α which will split apart the bifurcations corresponding to $\mathbf{R}^2$ (vortices) and $\mathbf{R}^2 \otimes \mathbf{C}$ (spiral cells and Z_2^2).

We choose to do this by replacing P^1 in (5.2) by $\alpha + P^1$. If $\alpha < 0$, then the bifurcation to vortices occurs second (in the bifurcation parameter $\lambda = \Omega_i$); if $\alpha > 0$, it occurs first. Thus, we may think of α as $\Omega_0 - \Omega_0^*$. See Fig. 6.1.

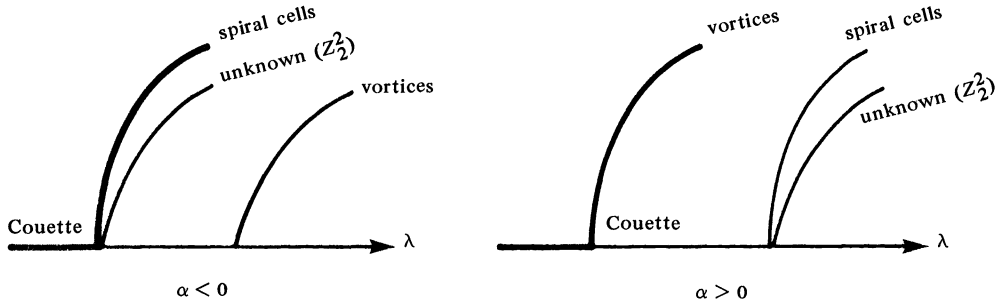


FIG. 6.1. *Schematic rendition of the effect of the perturbation α. The directions of the branches are chosen arbitrarily and secondary branches have been suppressed.*

The nondegeneracy conditions we impose on h are stated when $\alpha = 0$ and when $z = 0, A = 0$.

In Table 6.1 we list sixteen nondegeneracy conditions; a Γ-equivariant bifurcation

TABLE 6.1
Nondegeneracy conditions for h.

(a)	p_β	(h)	R^2
(b)	p_λ	(i)	$2P^1_N + R^2$
(c_1)	$(P^1_\beta + Q^3)p_\lambda - P^1_\lambda p_\beta$	(j)	$2(p_N P^1_\lambda - P^1_{Np\lambda}) - p_\lambda R^2$
(c_2)	$(P^1_\beta - Q^3)p_\lambda - P^1_\lambda p_\beta$	(k)	$P^1_N p_\beta - (P^1_\beta + Q^3)(p_N + r)$
(d)	P^1_N	(l)	$P^1_N p_\beta - (P^1_\beta - Q^3)(p_N - r)$
(e)	P^1_λ	(m)	r
(f)	$P^1_\lambda(p_N + r) - p_\lambda P^1_N$	(n)	Q^3
(g)	$P^1_\lambda(p_N - r) - p_\lambda P^1_N$	(o)	q

problem h is called *nondegenerate* if these sixteen expressions are nonzero for h. In Table 6.2 we list the lower order terms for each solution branch of $h=0$ and each eigenvalue of dh along these branches. For nondegenerate h these lower order terms determine the direction of branching (super or subcritical) and the (orbital) asymptotic stability of each solution. In Table 6.1 we use the convention that all quantities are to be evaluated at the origin. So, for example, P_N^1 means $P_N^1(0,0,0,0,0,0)$. The terms (a)–(n) in Table 6.2 refer to the corresponding expressions in Table 6.1.

TABLE 6.2

Branching equations and eigenvalues to lowest order for nondegenerate problems.

	Branching equations	Sign of real part of eigenvalues	Multiplicity
Couette flow $(O(2)\times S^1)$ $(0,0)$	None	$(b)\lambda$ $\alpha+(e)\lambda$	2 4
Taylor vortices $Z_2(\kappa)\times S^1$ $(x,0)$	$\lambda=-\dfrac{(a)}{(b)}x^2$	0 (a) $\alpha+\dfrac{(c_1)}{(b)}x^2$ $\alpha+\dfrac{(c_2)}{(b)}x^2$	1 1 2 2
unknown Z_2^2 $\left(0,\begin{pmatrix} a & 0\\ 0 & 0\end{pmatrix}\right)$	$\lambda=\dfrac{-1}{(e)}[\alpha+(d)a^2]$	0 $-\dfrac{(b)}{(e)}\alpha+\dfrac{(f)}{(e)}a^2$ $-\dfrac{(b)}{(e)}\alpha+\dfrac{(g)}{(e)}a^2$ (d) (h)	2 1 1 1 1
spiral cells $\widetilde{SO}(2)$ $\left(0,\begin{pmatrix} a & 0\\ 0 & a\end{pmatrix}\right)$	$\lambda=-\dfrac{1}{(e)}[\alpha+(i)a^2]$	0 $-\dfrac{(b)}{(e)}\alpha+\dfrac{(j)}{(e)}a^2$ (i) $-(h)$	1 2 1 2
wavy vortices $Z_2(\kappa\pi,\pi)$ $\left(iy,\begin{pmatrix} a & 0\\ 0 & 0\end{pmatrix}\right)$	$-\dfrac{(c_2)}{(b)}y^2+\dfrac{(g)}{(b)}a^2=\alpha$ $\lambda=\dfrac{(a)}{(c_2)}\alpha+\dfrac{(l)}{(c_2)}a^2$	0 (m) $2(n)y^2+(h)a^2$ $\det\begin{bmatrix} Y_y & Y_a\\ A_y & A_a\end{bmatrix}=(l)$ $\operatorname{tr}\begin{bmatrix} Y_y & Y_a\\ A_y & A_a\end{bmatrix}=(a)y^2+(d)a^2$	2 1 1 2

TABLE 6.2 (continued)

twisted vortices $Z_2(\kappa)$ $\left(x,\begin{pmatrix} a & 0 \\ 0 & 0 \end{pmatrix}\right)$	$-\frac{(c_1)}{(b)}x^2+\frac{(f)}{(b)}a^2=\alpha$ $\lambda=\frac{(a)}{(c_1)}\alpha+\frac{(k)}{(c_1)}a^2$	0 $-(m)$ $2(n)x^2+(h)a^2$ $\det\begin{bmatrix} X_x & X_a \\ A_x & A_a \end{bmatrix}=(k)$ $\operatorname{tr}\begin{bmatrix} X_x & X_a \\ A_x & A_a \end{bmatrix}=(a)x^2+(d)a^2$	2 1 1 2
$Z_2(\pi,\pi)$	No solutions by (h)		
1	No solutions by (m), (o) except perhaps on the orbit	$\left(x+iy,\begin{pmatrix} a & 0 \\ 0 & d \end{pmatrix}\right)$	

*The terms (a)–(o) are defined in Table 6.1 and required, by assumption of nondegeneracy, to be nonzero.

In our analysis of the bifurcation diagrams and the asymptotic stability of the associated solutions, we use only the equivariant form of the bifurcation equations and the implicit function theorem. Moreover, in each appeal to the implicit function theorem we find a neighborhood of the origin in (z,A,λ,α)-space on which its consequences are valid. Since we use the implicit function theorem only finitely many times, all of our conclusions hold simultaneously in some fixed neighborhood of $(0,0,0,0)$ in (z,A,λ,α)-space. This neighborhood does depend, however, on the particular values that enter into the nondegeneracy conditions.

The computation of the entries in Table 6.2 may be completed in a routine fashion using the entries in Table 5.1. We give the flavor of these computations by presenting the results for wavy vortices.

The branching equations for $Z_2(\kappa\pi,\pi)$ are

$$p(y^2,a,0^2,-a^2y^2,0,\lambda)-a^2r(y^2,a^2,0,-a^2y^2,0,\lambda)=0,$$
$$\alpha+P(y^2,a^2,0,a^2y^2,0,\lambda)+y^2Q^3(y^2,a^2,0,-a^2y^2,0,\lambda)=0.$$

See Table 5.1. (The third branching equation in that table is used only to eliminate τ.) Expanding to lowest order, we have

$$0=p_\beta(0)y^2+(p_N(0)-r(0))a^2+p_\lambda(0)\lambda+\cdots,$$

$$0=\alpha+P_\beta(0)y^2+P_N(0)a^2+P_\lambda(0)\lambda-Q^3(0)y^2+\cdots.$$

Using the implicit function theorem, we can solve for λ and y^2 as a function of a^2 if

$$p_\beta(0)P^1_\lambda(0)-p_\lambda(0)\left(P^1_\beta(0)-Q^3(0)\right)\neq 0.$$

This is condition (c_2) of Table 6.1. We then obtain

$$\lambda = \frac{p_\beta(0)}{(c_2)}\alpha + \frac{P_N^1(0)p_\beta(0) - \big(P_\beta^1(0) - Q^3(0)\big)\big(p_N(0) - r(0)\big)}{(c_2)}a^2 + \cdots,$$

$$y^2 = -\frac{p_\lambda(0)}{(c_2)}\alpha - \frac{p_\lambda(0)p_N^1(0) - P_\lambda^1(0)\big(p_N^1(0) - r(0)\big)}{(c_2)}a^2 + \cdots.$$

Using the entries in Table 6.1 along with some rearrangement of terms, we obtain the entry in Table 6.2.

7. Comparison with experiment. In this section we discuss how the above model bifurcation problem(s) on the six-dimensional kernel compare with experimental observations in the two main settings.

(1) Experiments by Andereck, Liu, and Swinney [1984] in the counterrotating case, including parameter values near a point at which the six-dimensional kernel appears to occur.
(2) The standard "main sequence" of bifurcations in the case where the outer cylinder is held fixed:

Couette flow → Taylor vortices → wavy vortices → modulated wavy vortices → ···

We will show below that it is possible to make choices for the signs of the coefficients that appear in Table 6.1 as nondegeneracy conditions, so that the resulting bifurcation sequences are in qualitative agreement with the experimentally observed bifurcation sequences. In the counterrotating case we have direct (numerical) evidence for the existence of the six-dimensional kernel through the work of DiPrima and Grannick [1971]; no evidence for this six-dimensional kernel currently exists when the outer cylinder is held fixed. We hasten to add, however, that the existence of the six-dimensional kernel is, because of symmetry, only a codimension one phenomenon; it should occur frequently in various forms of Taylor–Couette flow. We also note that, unfortunately, there are many different choices for the signs of the nondegeneracy conditions in Table 6.1 (over 10,000), so many different bifurcation sequences are possible besides the ones we consider here. However, the possibilities are not totally arbitrary, as we see below.

7.1. The counterrotating case. In a private communication, D. Andereck gave us the (qualitative) form of the experimental results for counterrotating Taylor–Couette flow, which have since appeared in Andereck, Liu, and Swinney [1984]. We present these results in Fig. 7.1. There are three features that deserve mention here.

(i) There is a critical speed of counterrotation Ω_0^* at which the primary bifurcation from Couette flow changes from Taylor vortices to spiral cells. This corresponds to the critical speed of counterrotation found numerically by DiPrima and Grannick [1971]. However, they presumably performed the calculations for values of the dimensions of the apparatus which differ from those used by Andereck, Liu and Swinney [1984].

(ii) In the weakly counterrotating case $\Omega_0 < \Omega_2^*$—this corresponds to the case

($\alpha > 0$) where our perturbation parameter α is positive—the observed bifurcation sequence is:

Couette → vortices → wavy vortices

See Fig. 7.2.

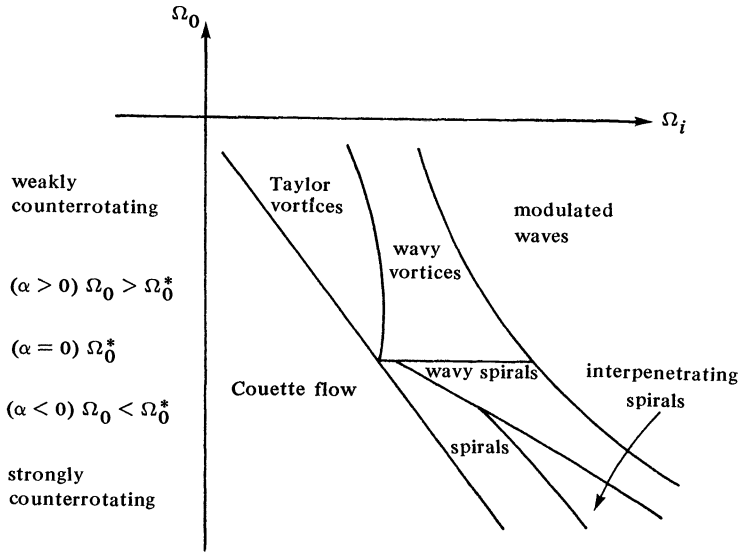


FIG. 7.1. *Qualitative version of the experimental results of Andereck, Liu, and Swinney* [1984] *showing observed transitions between stable states in the counterrotating Taylor–Couette system.*

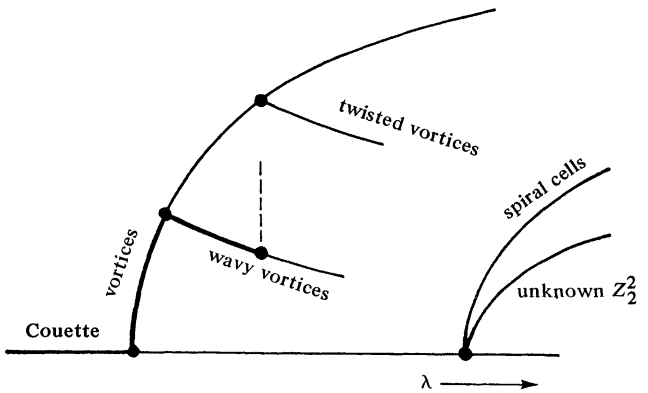


FIG. 7.2. *Schematic bifurcation diagram when* $\alpha > 0$ *corresponding to the observations of Andereck, Liu, and Swinney* [1984]. (*Secondary branches may or may not join other branches, depending on the values of the coefficients.*)

If the speed of the inner cylinder is increased further, then the wavy vortices lose stability to another state which does not appear to correspond to any state in our model.

(iii) In the strongly counterrotating case where Ω_0 is slightly greater than Ω_0^*—this corresponds to our $\alpha < 0$—the observed transition sequence is:

Couette → spiral cells → wavy spiral cells

See Fig. 7.3.

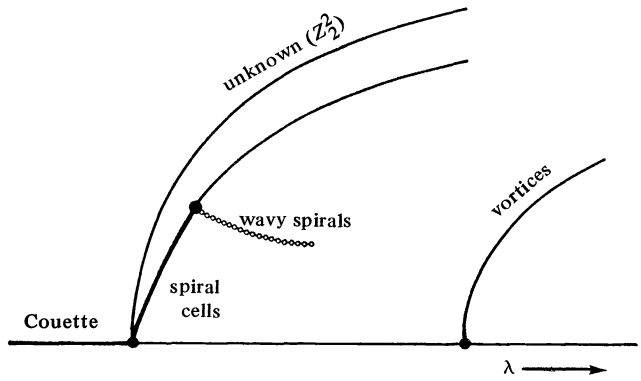


FIG. 7.3. *Schematic bifurcation diagram when* $\alpha < 0$ *corresponding to the observations of Andereck* [1984].

In order to reproduce qualitatively the experimental results, we demand the following:

(7.1)
(a) Couette flow is stable for $\lambda \ll 0$.
(b) Vortices bifurcate supercritically and stably when $\alpha > 0$.
(c) There is a secondary bifurcation from vortices to wavy vortices when $\alpha > 0$.
(d) Wavy vortices are supercritical and stable at the initial bifurcation from vortices.
(e) Spiral cells bifurcate supercritically and stably when $\alpha < 0$.
(f) Spiral cells lose stability to a Hopf bifurcation.

We claim that it is possible to choose signs for the nondegeneracy conditions in Table 6.1 so that each of the conditions in (7.1) is satisfied. Moreover, we claim that when these nondegeneracy conditions are satisfied, it follows that:

(7.2) Any bifurcation from vortices to Taylor vortices occurs after the bifurcation from vortices to wavy vortices.

The assumptions in (7.1) correspond, in order, to the following nondegeneracy conditions (cf. Table 6.1):

(7.3)
(a) $(e) < 0, (b) < 0$.
(b) $(a) > 0$.
(c) $(c_2) > 0$.
(d) $(l) > 0; (m) > 0, (n) > 0$.
(e) $(i) > 0, (h) < 0$.
(f) $(j) > 0$.

It is a simple matter to check that the conditions (7.3) are precisely the conditions needed to satisfy (7.1). The only point which requires comment is asymptotic stability of the wavy vortex branch. Observe that at the bifurcation from vortices to wavy vortices, $a = 0$ and $y \neq 0$. It follows from Table 6.2 that the wavy vortices are stable if

$$(m) > 0, \quad (n) > 0, \quad (l) > 0, \quad \text{and } (a) > 0.$$

However, $(l) > 0$ when the branch of wavy vortices is supercritical, and $(a) > 0$ has already been assumed in (7.3b). Observe that the branch of wavy vortices can lose

stability if either

$$(d)<0 \quad \text{or} \quad (h)<0. \tag{7.4}$$

If $(d)<0$ then this branch will lose stability to a torus bifurcation. However,

$$(d)=((i)-(h))/2, \tag{7.5}$$

and the assumption that spiral cells are stable implies that $(i)>0$, $(h)<0$ (cf. (7.3e)) so this possibility cannot occur. Nevertheless, $(h)<0$ is satisfied, and wavy vortices may lose stability by a single real eigenvalue passing through zero. This observation verifies (7.2a). It is possible that a new solution branch with isotropy subgroup 1 will appear at this bifurcation, but we have neither confirmed nor eliminated this possibility. See Table 5.1.

Finally, we verify (7.2b). The wavy vortex branch begins at $\lambda_w(a)/(c_2)$, while a branch of twisted vortices would begin at $\lambda_t=(a)/(c_2)$. We compute $\operatorname{sgn}(\lambda_t-\lambda_w)$. Now

$$\operatorname{sgn}(\lambda_t-\lambda_w)=\operatorname{sgn}\left(\frac{1}{(c_1)}-\frac{1}{(c_2)}\right), \tag{7.6}$$

since $(a)>0$ by (7.3b). However,

$$(c_2)-(c_1)=-2Q^3p_\lambda=-2(n)(b)>0$$

using (7.3a, d). Hence (7.6) implies that $\lambda_t>\lambda_w$ as claimed in (7.2b). Observe that it is possible, under different circumstances, for twisted vortices to bifurcate supercritically and stably from vortices. Such a transition has been observed in the corotating case. See Andereck, Dickman and Swinney [1983].

7.2. The main sequence. Here we verify that the main sequence of bifurcations can also occur in the six-dimensional model. This sequence of bifurcations is observed in experiments when the outer cylinder is held fixed. For this sequence to hold, we need Couette flow to lose stability first to vortices. This happens in our model when $\alpha>0$, and we concentrate on this case.

To obtain the main sequence, we need (7.1a, b, c, d) to hold. Of course, this is possible precisely when the nondegeneracy conditions (7.3a, b, c, d) hold. If we wish to show in this model that, in addition, the wavy vortex solutions lose stability to a torus bifurcation, then two complex conjugate eigenvalues must cross the imaginary axis along the branch of wavy vortices. This can happen only if $(d)<0$. Note that if $(d)<0$ then (7.5) implies that (7.3e) is not valid, and that spiral cells cannot be asymptotically stable.

As we saw above, wavy vortices can lose stability by a real eigenvalue crossing through 0. However, this eventuality cannot occur if $(h)>0$, which is possible, since we have assumed nothing about (i). Thus, assuming

$$(d)<0, \qquad (h)>0$$

leads to the main sequence. (See Fig. 7.4.) Other choices for the main sequence are possible.

We conclude that both the main sequence and certain regimes in the experiments of Andereck, Liu, and Swinney [1984] appear to be qualitatively consistent with our six-dimensional model, for suitable values of the coefficients. (Note that aside from the states discussed above, no other stable states occur except perhaps with isotropy group

1, as mentioned.) Since the coefficients can in principle be computed by Lyapunov–Schmidt reduction from the Navier–Stokes equations, further numerical work should be able to provide a more stringent test. It would also be of interest to determine, in terms of the physical parameters in the problem, the location of the codimension one set of values at which the six-dimensional kernel occurs.

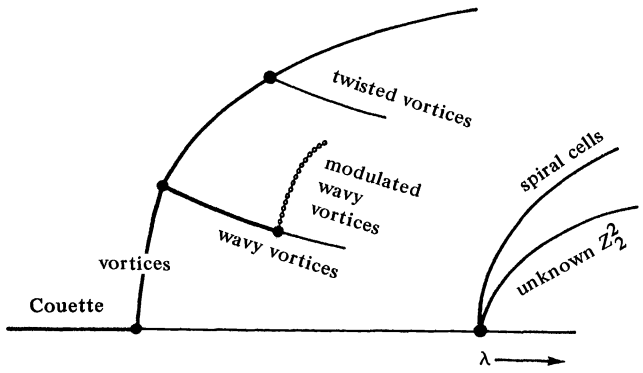


FIG. 7.4. *Schematic bifurcation diagram corresponding to (one occurrence of) the main sequence.*

Appendix. Equivariant mappings on the six-dimensional kernel. Let $V=\mathbf{R}^2\oplus(\mathbf{R}^2\otimes\mathbf{C})$ be the six-dimensional kernel described in §4. Recall that $\Gamma=O(2)\times S^1$ acts on V by

$$(\theta,\psi)(v,w\otimes z)=\big(R_\theta v,(R_\theta w)\otimes(e^{i\psi}Z)\big).$$

For computational purposes we choose coordinates by identifying the first $\mathbf{R}^2$ with $\mathbf{C}$, and $\mathbf{R}^2\otimes\mathbf{C}$ with 2×2 matrices as described in §4, so that an element of V is written (z,a) where

$$z\in\mathbf{C},\quad A=\begin{pmatrix}a & b\\ c & d\end{pmatrix},\quad a,b,c,d\in\mathbf{R}.$$

Recall that a (smooth) function $\phi\colon V\to\mathbf{R}$ is *invariant* under Γ if

$$\phi(\gamma v)=\phi(v),\qquad \gamma\in\Gamma,\quad v\in V,$$

and a (smooth) mapping $\Phi\colon V\to V$ is *equivariant* if it commutes with Γ, that is

$$\Phi(\gamma v)=\gamma\Phi(v),\qquad \gamma\in\Gamma,\quad v\in V.$$

The aim of this appendix is to describe completely these invariant functions and equivariant mappings, as promised in §4 above. The main result, which will yield Theorem 4.3 when appropriate terms are collected together, is:

PROPOSITION A.1. (a) *Every invariant function on V is of the form*

$$\phi(v)=h(\beta,N,\delta^2,\gamma,\sigma) \tag{A.1}$$

for a smooth function $h\colon\mathbf{R}^5\to\mathbf{R}$, where

$$\begin{aligned}\beta&=z\bar z=x^2+y^2,\\ N&=a^2+b^2+c^2+d^2,\\ \delta^2&=(ad-bc)^2,\\ \gamma&=\mathrm{Re}(z^2\bar\zeta),\\ \sigma&=\delta\,\mathrm{Im}(z^2\bar\zeta)\end{aligned} \tag{A.2}$$

and

$$\zeta = (a^2+b^2-c^2-d^2)+2i(ac+bd). \tag{A.3}$$

(b) *Every equivariant mapping $V \to V$ is of the form*

$$\begin{aligned}\Phi(z,A) = \big(& pz + qi\delta z + r\bar{z}\zeta + si\delta\bar{z}\zeta, \\ & P^1 \operatorname{Re}(H_1) + P^2 \operatorname{Re}(H_2) \\ & + Q^1 \operatorname{Im}(z^2\bar{\zeta}) \operatorname{Im}(H_1) + Q^2 \operatorname{Im}(z^2\bar{\zeta}) \operatorname{Im}(H_2) \\ & + Q^3 \operatorname{Re}(\bar{z}^2 H_3) + Q^4 \operatorname{Re}(\bar{z}^2 H_4) \\ & + R^1\delta \operatorname{Im}(H_1) + R^2\delta \operatorname{Im}(H_2) + R^3\delta \operatorname{Im}(\bar{\zeta}H_3) + R^4\delta \operatorname{Im}(\bar{\zeta}H_4) \\ & + M^3\delta \operatorname{Im}(\bar{z}^2 H_3) + M^4\delta \operatorname{Im}(\bar{z}^2 H_4)\big),\end{aligned} \tag{A.4}$$

where

$$H_1 = \begin{pmatrix} a-ic & b-id \\ c+ia & d+ib \end{pmatrix}, \qquad H_3 = \begin{pmatrix} a+ic & b+id \\ -c+ia & -d+ib \end{pmatrix},$$
$$H_2 = \begin{pmatrix} -b+id & a-ic \\ -d-ib & c+ia \end{pmatrix}, \qquad H_4 = \begin{pmatrix} -b-id & a+ic \\ d-ib & -c+ia \end{pmatrix},$$

and

$$p, q, r, s, P^1, P^2, Q^1, Q^2, Q^3, Q^4, R^1, R^2, R^3, R^4, M^3, M^4$$

are invariant functions.

In more abstract language, Proposition A.1 says that the ring of invariant functions is *generated* by $\beta, N, \delta^2, \gamma, \sigma$; and that the module of equivariant mappings is generated over the invariants by the twelve mappings in (A.4), whose coefficients are $p, q, r, s, P^1, \cdots, M^4$. By standard results of Schwarz [1975] and Poénaru [1976] we may assume ϕ and Φ are polynomials when proving Proposition A.1.

The computation comes in two stages. First we compute the (polynomial) S^1-invariants and -equivariants; then we use this information and the $O(2)$-action to obtain the $O(2)\times S^1$-invariants and -equivariants. Since S^1 acts trivially on $z \in \mathbf{R}^2$, we need consider only the action on $A \in \mathbf{R}^2 \otimes \mathbf{C}$. We take complex coordinates

$$z_1 = a+ib, \qquad z_2 = c+id.$$

Then we may identify $\mathbf{R}^2 \otimes \mathbf{C}$ with $\mathbf{C} \oplus \mathbf{C}$, where S^1 acts diagonally:

$$\theta(z_1, z_2) = (e^{i\theta}z_1, e^{i\theta}z_2).$$

LEMMA A.2. *The real S^1-invariants on $\mathbf{C}\oplus\mathbf{C}$ are generated by $z_1\bar{z}_1$, $z_2\bar{z}_2$, $\operatorname{Re} z_1\bar{z}_2$, $\operatorname{Im} z_1\bar{z}_2$. The S^1-equivariants are generated over the invariants by $(z_1, 0)$, $(0, z_1)$, $(z_2, 0)$, $(0, z_2)$, $(i\bar{z}_1, 0)$, $(0, i\bar{z}_1)$, $(i\bar{z}_2, 0)$, $(0, i\bar{z}_2)$.*

Proof. These results (which generalize easily to S^1 acting on $\mathbf{C}^n$) are no doubt well-known, but for completeness we sketch a proof. The idea is first to find the complex invariants and equivariants and then to read off the real ones.

Consider a $\mathbf{C}$-valued polynomial function

$$p(z_1, \bar{z}_1, z_2, \bar{z}_2) = \sum A_{\alpha\beta\gamma\delta} z_1^{\alpha} \bar{z}_1^{\beta} z_2^{\gamma} \bar{z}_2^{\delta}.$$

Since $\overline{e^{i\theta}z} = e^{-i\theta}\bar{z}$, we can use S^1-invariance to exclude all terms other than those for which

$$\alpha - \beta + \gamma - \delta = 0.$$

So p is a polynomial in $z_1\bar{z}_1$, $z_2\bar{z}_2$, $\bar{z}_1 z_2$, and $z_1\bar{z}_2$. If p is to be real in a, b, c, d, then we have $p = \bar{p}$, so $A_{\alpha\beta\gamma\delta} = \bar{A}_{\beta\alpha\delta\gamma}$. This leads to the real invariant generators stated.

For the equivariants, we consider a pair of functions p_1, p_2 of the above form. Equivariance excludes all terms other than those for which

$$\alpha - \beta + \gamma - \delta = 1.$$

This yields equivariant generators which are *complex* scalar multiples of $(z_1, 0)$, $(z_2, 0)$, $(0, z_1)$, $(0, z_2)$. Taking real and imaginary parts, we obtain the stated real equivariant generators.

In $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ coordinates, we have the invariant generators

$$\begin{aligned} &z_1\bar{z}_1 = a^2 + b^2, \\ &z_2\bar{z}_2 = c^2 + d^2, \\ &\mathrm{Re}(z_1\bar{z}_2) = ac + bd, \\ &\mathrm{Im}(z_1\bar{z}_2) = bc - ad = -\delta, \end{aligned} \tag{A.5}$$

and the equivariant generators $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \to$

$$E_1 = \begin{pmatrix} a & b \\ 0 & 0 \end{pmatrix}, \quad E_2 = \begin{pmatrix} 0 & 0 \\ a & b \end{pmatrix}, \quad E_3 = \begin{pmatrix} 0 & 0 \\ c & d \end{pmatrix}, \quad E_4 = \begin{pmatrix} c & d \\ 0 & 0 \end{pmatrix}, \tag{A.6}$$

$$E_5 = \begin{pmatrix} -b & a \\ 0 & 0 \end{pmatrix}, \quad E_6 = \begin{pmatrix} 0 & 0 \\ -b & a \end{pmatrix} \quad E_7 = \begin{pmatrix} 0 & 0 \\ -d & c \end{pmatrix}, \quad E_8 = \begin{pmatrix} -d & c \\ 0 & 0 \end{pmatrix}.$$

Note that there is a relation

$$(a^2 + b^2)(c^2 + d^2) = (ac + bd)^2 + (ad - bc)^2.$$

We are now ready for the:

Proof of Proposition A.1 (a). The calculations are easier if we use complex notation. Let

$$\zeta = (a^2 + b^2 - c^2 - d^2) + 2i(ac + bd).$$

Then every real-valued function of the four invariant generators can be written in terms of $N = a^2 + b^2 + c^2 + d^2$, ζ, $\bar{\zeta}$, and δ. Note that N and δ^2 are $O(2)$-invariant. See Golubitsky and Stewart [1985, §9].

Since S^1 acts trivially on $z \in \mathbf{R}^2$, we can write the S^1-invariants on $\mathbf{R}^2 \oplus (\mathbf{R}^2 \otimes \mathbf{C})$ in the form

$$\phi(z, \bar{z}, N, \zeta, \bar{\zeta}, \delta).$$

Under the $O(2)$-action these transform as follows:

	z	$\bar{z}$	ζ	$\bar{\zeta}$	δ	N
κ	$\bar{z}$	z	$\bar{\zeta}$	ζ	$-\delta$	N
θ	$e^{i\theta}z$	$e^{-i\theta}\bar{z}$	$e^{2i\theta}\zeta$	$e^{-2i\theta}\bar{\zeta}$	δ	N

(The expressions ζ and $\bar{\zeta}$ are introduced because of this pleasant transformation behavior.)

Since $z\bar{z}$, $\zeta\bar{\zeta}$, N, and δ^2 are $O(2)$-invariant, we can write the general $O(2)\times S^1$-invariant in the form

$$\begin{aligned}\phi &= az^\alpha\zeta^\beta + bz^\alpha\bar{\zeta}^\beta + c\bar{z}^\alpha\zeta^\beta + d\bar{z}^\alpha\bar{\zeta}^\beta + \delta\left(ez^\alpha\zeta^\beta + fz^\alpha\bar{\zeta}^\beta + g\bar{z}^\alpha\zeta^\beta + h\bar{z}^\alpha\bar{\zeta}^\beta\right)\\ &\equiv \phi_0 + \delta\phi_1\end{aligned} \tag{A.7}$$

where $a,b,\cdots,h\in\mathbf{C}[z\bar{z},N,\delta^2]$. (Note: $\zeta\bar{\zeta}=N-4\delta^2$ so no $\zeta\bar{\zeta}$ terms are required.) Reality of ϕ implies that

$$\bar{a}=d,\quad \bar{b}=c,\quad \bar{e}=h,\quad \bar{f}=g,$$

while κ-invariance leads to

$$a=b,\quad c=d,\quad e=-h,\quad f=-g.$$

Hence a,b,c,d are real and e,f,g,h are purely imaginary.

Finally we apply $SO(2)$-invariance. Since δ is $SO(2)$-invariant and is independent of $z,\bar{z},\zeta,\bar{\zeta}$, we must have ϕ_0 and ϕ_1 separately $SO(2)$-invariant. This excludes all terms other than

$$a\left(z^\alpha\zeta^\beta+\bar{z}^\alpha\bar{\zeta}^\beta\right) \quad \text{when } \alpha+2\beta=0, \tag{A.8}$$

$$b\left(z^\alpha\bar{\zeta}^\beta+\bar{z}^\alpha\zeta^\beta\right) \quad \text{when } \alpha-2\beta=0, \tag{A.9}$$

$$\delta e\left(z^\alpha\zeta^\beta-\bar{z}^\alpha\bar{\zeta}^\beta\right) \quad \text{when } \alpha+2\beta=0, \tag{A.10}$$

$$\delta f\left(z^\alpha\bar{\zeta}^\beta-\bar{z}^\alpha\zeta^\beta\right) \quad \text{when } \alpha-2\beta=0, \tag{A.11}$$

Now (A.8) and (A.10) imply $\alpha=\beta=0$, giving nothing new. The others yield $\alpha=2\beta$.

We claim that only $\alpha=2$, $\beta=1$ yield new generators. For example

$$\left(z^{\alpha+2}\bar{\zeta}^{\beta+1}+\bar{z}^{\alpha+2}\zeta^{\beta+1}\right)=\left(z^\alpha\bar{\zeta}^\beta+\bar{z}^\alpha\zeta^\beta\right)\left(z^2\bar{\zeta}+\bar{z}^2\zeta\right)-\left(z\bar{z}\right)^2\left(\zeta\bar{\zeta}\right)\left(z^{\alpha-2}\bar{\zeta}^{\beta-1}+\bar{z}^{\alpha-2}\zeta^{\beta-1}\right).$$

Since b is real and f purely imaginary, we obtain generators

$$\mathrm{Re}\left(\bar{z}^2\bar{\zeta}\right),\ i\delta\,\mathrm{Im}\left(z^2\bar{\zeta}\right)$$

in addition to $z\bar{z}$, N, δ^2. This proves part (a) of Proposition A.1.

Proof of Proposition A.1 (b). Write the general equivariant in the form

$$\Phi(z,A)=\left(\Phi_0(z,A),\Phi_1(z,A)\right)$$

where

$$\Phi_0\colon \mathbf{R}^2\oplus\left(\mathbf{R}^2\otimes\mathbf{C}\right)\to\mathbf{R}^2,$$
$$\Phi_1\colon \mathbf{R}^2\oplus\left(\mathbf{R}^2\otimes\mathbf{C}\right)\to\mathbf{R}^2\otimes\mathbf{C}.$$

We begin with Φ_0. Since the S^1-action on $\mathbf{R}^2$ is *trivial*, the S^1-equivariance condition implies that Φ_0 is S^1-invariant and hence can be written in the form (A.7) above.

However, this time there is no reality condition since we seek mappings into $\mathbf{R}^2$, not $\mathbf{R}$. The κ-equivariance again implies a,b,c,d are real, and e,f,g,h are purely imaginary. Replacing the latter by ie, if, ig, ih, we may assume all coefficients $a-h$ are real, and replace δ by $i\delta$. Write $\Phi_0=\phi_0+i\delta\phi_1$: again we can treat ϕ_0 and ϕ_1 separately.

Now $SO(2)$-equivariance excludes all terms other than

$$z^\alpha\zeta^\beta,\ i\delta z^\alpha\zeta^\beta;\quad \alpha+2\beta=1, \tag{A.12}$$

$$z^\alpha\bar{\zeta}^\beta,\ i\delta z^\alpha\bar{\zeta}^\beta;\quad \alpha-2\beta=1, \tag{A.13}$$

(A.14) $$\bar{z}^{\alpha}\zeta^{\beta},\ i\delta\bar{\zeta}^{\alpha}\zeta^{\beta};\quad -\alpha+2\beta=1,$$

(A.15) $$\bar{z}^{\alpha}\bar{\zeta}^{\beta},\ i\delta\bar{z}^{\alpha}\bar{\zeta}^{\beta};\quad -\alpha-2\beta=1.$$

In (A.12) we have $\alpha=1$, $\beta=0$, yielding z and $i\delta z$. In (A.13) we have $\alpha=2\beta+1$. As before, we may use the invariance of $z\bar{z}$ and $\zeta\bar{\zeta}$ to reduce the size of α and β:

$$\begin{aligned} z^{2\beta+1}\bar{\zeta}^{\beta} &= z^{2\beta-1}\left(z^2\bar{\zeta}\right)\bar{\zeta}^{\beta-1} \\ &= \left(z^2\bar{\zeta}+\bar{z}^2\zeta\right)z^{2\beta-1}\bar{\zeta}^{\beta-1}-\left(z\bar{z}\right)^2\left(\zeta\bar{\zeta}\right)z^{2\beta-3}\bar{\zeta}^{\beta-2}. \end{aligned}$$

Thus we can reduce β by 1 and α by 2. The process stops when $\beta=1$, $\alpha=2$. But now

$$z^3\bar{\zeta}=zz^2\bar{\zeta}=\left(z^2\bar{\zeta}+\bar{z}^2\zeta\right)z-\left(z\bar{z}\right)\bar{z}\zeta.$$

Thus we get new generators $\bar{z}\zeta$, $i\delta\bar{z}\zeta$. In (A.14) we can similarly assume $\beta\leqq 2$. But $\beta=2$ gives

$$\bar{z}^3\zeta^2=\left(\bar{z}^2\zeta+z^2\bar{\zeta}\right)\bar{z}\zeta-\left(z\bar{z}\right)\left(\zeta\bar{\zeta}\right)z$$

so no new generator arises; and $\beta=1$ gives $\bar{z}\zeta$ which is already included. Finally (A.15) is not possible.

Thus we have found four generators $(z,0)(i\delta z,0)$, $(\bar{z}\zeta,0)$, $(i\delta\bar{z}\zeta,0)$ corresponding to mappings of $\mathbf{R}^2\oplus(\mathbf{R}^2\otimes\mathbf{C})$ into $\mathbf{R}^2$.

Now we look at

$$\Phi_1\colon \mathbf{R}^2\oplus(\mathbf{R}^2\otimes\mathbf{C})\to\mathbf{R}^2\otimes\mathbf{C}.$$

Again complex notation is more convenient. Define (in a notation consistent with the statement of Proposition A.1) the complex matrices

(A.16)
$$\begin{aligned} H_1&=(E_1+E_3)+i(E_2-E_4),\\ H_2&=(E_5+E_7)+i(E_6-E_8),\\ H_3&=(E_1-E_3)+i(E_2+E_4),\\ H_4&=(E_5-E_7)+i(E_6+E_8). \end{aligned}$$

Then the S^1-equivariants on $\mathbf{R}^2\otimes\mathbf{C}$ are generated over $\mathbf{C}$ by H_k, $\overline{H}_k$ $(k=1,\cdots,4)$ and over $\mathbf{R}$ by the real and imaginary parts of H_k $(k=1,\cdots,4)$. Since S^1 acts trivially on $z\in\mathbf{R}^2$ we can think of the S^1-equivariants mapping $\mathbf{R}^2\oplus(\mathbf{R}^2\otimes\mathbf{C})\to\mathbf{R}^2$ as S^1-equivariants mapping $\mathbf{R}^2\otimes\mathbf{C}\to\mathbf{R}^2$ parametrized by z and $\bar{z}$. Thus they are linear combinations of H_k, $\overline{H}_k$, $(k=1,\cdots,4)$ with coefficients in $\mathbf{C}[N,\delta,\zeta,\bar{\zeta};z,\bar{z}]$.

We write the equivariance condition $\phi(\gamma v)=\gamma\Phi(v)$ in the form

(A.17) $$\Phi(v)=\gamma^{-1}\Phi(\gamma v).$$

Suppose

$$\Phi(v)=\rho(v)H(v)$$

where $\rho\colon \mathbf{R}^2\oplus(\mathbf{R}^2\otimes\mathbf{C})\to\mathbf{R}$, $H\colon \mathbf{R}^2\oplus(\mathbf{R}^2\otimes\mathbf{C})\to\mathbf{R}^2\otimes\mathbf{C}$. Then (A.17) is equivalent to

(A.18) $$\rho(v)H(v)=\gamma^{-1}\rho(\gamma v)H(\gamma v)=\rho(\gamma v)\gamma^{-1}H(\gamma v).$$

We compute this action on H_k $(k=1,\cdots,4)$ when $\gamma\in O(2)$. Using (A.16) and noting that

$$\kappa(z,A)=\left(\bar{z},\begin{pmatrix}1&0\\0&-1\end{pmatrix}A\right),$$

$$\psi(z,A)=\left(e^{i\psi}z,R_\psi A\right)$$

where

$$R_\psi=\begin{pmatrix}\cos\psi&-\sin\psi\\\sin\psi&\cos\psi\end{pmatrix},$$

we find

(A.19)

$H(v)$	$\kappa^{-1}H(\kappa v)$	$\psi^{-1}H(\psi v)$
H_1	$\overline{H}_1$	H_1
H_2	$\overline{H}_2$	H_2
H_3	$\overline{H}_3$	$e^{2i\psi}H_3$
H_4	$\overline{H}_4$	$e^{2i\psi}H_4$

We can write the general S^1-equivariant Φ_1: $\mathbf{R}^2\oplus(\mathbf{R}\otimes\mathbf{C})\to\mathbf{R}^2\otimes\mathbf{C}$ in the form

(A.20) $$\Phi_1=\mathrm{Re}\left\{\sum_{k=1}^{4}(\rho_k+\delta\sigma_k)H_k\right\}$$

where the ρ_k, σ_k are polynomials over $\mathbf{C}$ of the form

$$\rho_k=\rho_k(N,\delta^2,\zeta,\bar{\zeta};z,\bar{z}),\qquad \sigma_k=\sigma_k(N,\delta^2,\zeta,\bar{\zeta};z,\bar{z}).$$

Since $z\bar{z}$ and $\zeta\bar{\zeta}$ are $O(2)$-invariant, we can write the ρ_k and σ_k as

(A.21) $$az^\alpha\zeta^\beta+bz^\alpha\bar{\zeta}^\beta+c\bar{z}^\alpha\zeta^\beta+d\bar{z}^\alpha\bar{\zeta}^\beta$$

with $a,b,c,d,\in\mathbf{C}[N,\delta^2]$.

We now apply κ- and ψ-equivariance in the form (A.17), writing

$$\tilde{\rho}=\rho(\kappa z,\kappa A),\qquad \hat{\rho}=\rho(\psi z,\psi A).$$

Now κ-equivariance (using (A.18) and (A.19)) implies that

(A.22) $$\tilde{\rho}_k=\bar{\rho}_k,$$

(A.23) $$\tilde{\sigma}_k=-\bar{\sigma}_k$$

and ψ-equivariance implies

(A.24)
(A.25) $$\hat{\rho}_k=\begin{cases}\rho_k & (k=1,2),\\ e^{-2i\psi}\rho_k & (k=3,4),\end{cases}$$

(A.26)
(A.27) $$\hat{\sigma}_k=\begin{cases}\sigma_k & (k=1,2),\\ e^{-2i\psi}\sigma_k & (k=3,4).\end{cases}$$

Now write the ρ_k and σ_k in the form (A.20) (we suppress unnecessary fine points of notation in the interests of clarity). From (A.22) we get

$$\text{(A.28)} \qquad (\text{for } \rho_k)\ a=\bar{a},\quad b=\bar{b},\quad c=\bar{c},\quad d=\bar{d},$$

and from (A.23)

$$\text{(A.29)} \qquad (\text{for } \sigma_k)\ a=-\bar{a},\quad b=-\bar{b},\quad c=-\bar{c},\quad d=-\bar{d}.$$

That is, the coefficients are real for ρ_k and purely imaginary for σ_k. We therefore replace σ_k by $i\sigma_k$, so that ρ_k and σ_k are real: now (A.20) takes the form

$$\text{(A.30)} \qquad \Phi=\operatorname{Re}\left\{\sum_{k=1}^{4}(\rho_k+i\delta\sigma_k)H_k\right\}.$$

The ψ-action multiplies z, ζ, and H_k by complex constants $e^{i\psi}$, $e^{2i\psi}$, $e^{2i\psi}$ respectively. Hence we may consider each of the eight terms in (A.30) separately. From (A.26) and (A.27) we obtain the following conditions on the exponents α, β, required for ψ-equivariance:

	Real part of:	$k=1,2$	$k=3,4$
(A.31)	$z^\alpha\zeta^\beta H_k$	$\alpha+2\beta=0$	$\alpha+2\beta=-2$
(A.32)	$z^\alpha\bar{\zeta}^\beta H_k$	$\alpha-2\beta=0$	$\alpha-2\beta=-2$
(A.33)	$\bar{z}^\alpha\zeta^\beta H_k$	$-\alpha+2\beta=0$	$-\alpha+2\beta=-2$
(A.34)	$\bar{z}^\alpha\bar{\zeta}^\beta H_k$	$-\alpha-2\beta=0$	$-\alpha-2\beta=-2$
(A.35)	$i\delta z^\alpha\zeta^\beta H_k$	$\alpha+2\beta=0$	$\alpha+2\beta=-2$
(A.36)	$i\delta z^\alpha\bar{\zeta}^\beta H_k$	$\alpha-2\beta=0$	$\alpha-2\beta=-2$
(A.37)	$i\delta\bar{z}^\alpha\zeta^\beta H_k$	$-\alpha+2\beta=0$	$-\alpha+2\beta=-2$
(A.38)	$i\delta\bar{z}^\alpha\bar{\zeta}^\beta H_k$	$-\alpha-2\beta=0$	$-\alpha-2\beta=-2$

We deal with these terms case by case, first for $k=1,2$; then for $k=3,4$. So let $k=1,2$.

(A.31) implies $\alpha=\beta=0$, leading to the generators

$$\text{(A.39)} \qquad \operatorname{Re}(H_k),\qquad k=1,2.$$

(A.32) requires $\alpha=2\beta$, so we get

$$\text{(A.40)} \qquad z^{2\beta}\bar{\zeta}^\beta H_k+\bar{z}^{2\beta}\zeta^\beta\overline{H}_k.$$

Similarly (A.33) requires $\alpha=2\beta$, and the result is

$$\text{(A.41)} \qquad \bar{z}^{2\beta}\zeta^\beta H_k+z^{2\beta}\bar{\zeta}^\beta\overline{H}_k.$$

Forming the sum and difference of (A.40) and (A.41) we may replace them by

$$x_\beta=(z^{2\beta}\bar{\zeta}^\beta+\bar{z}^{2\beta}\zeta^\beta)(H_k+\overline{H}_k),$$

$$y_\beta=(z^{2\beta}\bar{\zeta}^\beta-\bar{z}^{2\beta}\zeta^\beta)(H_k-\overline{H}_k).$$

We observe the following identities:

$$x_{\beta+1}=2\operatorname{Re}(z^2\bar{\zeta})x_\beta-(z^2\bar{z}^2\zeta\bar{\zeta})x_{\beta-1},$$

$$y_{\beta+1}=2\operatorname{Re}(z^2\bar{\zeta})y_\beta-(z^2\bar{z}^2\zeta\bar{\zeta})y_{\beta-1},$$

which have invariant functions as coefficients. Since

$$x_0 = 2\,\mathrm{Re}(H_k), \quad y_0 = 0, \quad x_1 = 2\,\mathrm{Re}(z^2\bar{\zeta})\mathrm{Re}(H_k),$$

and these may be obtained from the generators (A.39), an inductive argument shows that only y_1 need be retained in a list of generators. So we obtain the new generators

$$\mathrm{Im}(z^2\bar{\zeta})\mathrm{Im}(H_k), \qquad k = 1,2. \tag{A.42}$$

For (A.34) we have $\alpha = \beta = 0$ and nothing new results.

For (A.35) we have $\alpha = \beta = 0$, leading to new generators $\mathrm{Re}(i\delta H_k)$, or equivalently

$$\delta\,\mathrm{Im}(H_k), \quad k = 1,2. \tag{A.43}$$

From (A.36) and (A.37) we get $\alpha = 2\beta$, yielding

$$i\delta z^{2\beta}\bar{\zeta}^{\beta}H_k - i\delta\bar{z}^{2\beta}\zeta^{\beta}\overline{H}_k,$$
$$i\delta\bar{z}^{2\beta}\zeta^{\beta}H_k - i\delta z^{2\beta}\bar{\zeta}^{\beta}\overline{H}_k.$$

Forming the sum and difference, we replace these by

$$v_\beta = i\delta\left(z^{2\beta}\bar{\zeta}^{\beta} + \bar{z}^{2\beta}\zeta^{\beta}\right)\left(H_k - \overline{H}_k\right),$$
$$w_\beta = i\delta\left(z^{2\beta}\bar{\zeta}^{\beta} - \bar{z}^{2\beta}\zeta^{\beta}\right)\left(H_k + \overline{H}_k\right).$$

We note the identities

$$v_{\beta+1} = 2\,\mathrm{Re}\left(z^2\bar{\zeta}\right)v_\beta - \left(z^2\bar{z}^2\zeta\bar{\zeta}\right)v_{\beta-1},$$
$$w_{\beta+1} = 2\,\mathrm{Re}\left(z^2\bar{\zeta}\right)w_\beta - \left(z^2\bar{z}^2\zeta\bar{\zeta}\right)w_{\beta-1},$$
$$v_1 = \mathrm{Re}\left(z^2\bar{\zeta}\right)\delta\,\mathrm{Im}(H_k),$$
$$w_1 = \delta\,\mathrm{Im}\left(z^2\bar{\zeta}\right)\mathrm{Re}(H_k).$$

It follows by induction that no new generators arise here.

Finally (A.38) leads to $\alpha = \beta = 0$, and no new generators. This completes the analysis for $k = 1,2$.

Next, we let $k = 3,4$. The calculations follow a similar pattern.

(A.31) is impossible.

(A.32) and (A.33) lead to

$$t_\beta = z^{2\beta-2}\bar{\zeta}^{\beta}H_k + \bar{z}^{2\beta-2}\zeta^{\beta}\overline{H}_k, \qquad \beta \geqq 1,$$
$$u_\beta = \bar{z}^{2\beta+2}\zeta^{\beta}H_k + z^{2\beta+2}\bar{\zeta}^{\beta}H_k, \qquad \beta \geqq 0.$$

Now

$$t_{\beta+1} = 2\,\mathrm{Re}\left(z^2\bar{\zeta}\right)t_\beta - \left(z^2\bar{z}^2\zeta\bar{\zeta}\right)t_{\beta-1}, \qquad (\beta \geqq 2),$$
$$t_2 = 2\,\mathrm{Re}\left(z^2\bar{\zeta}\right)t_1 - \left(\zeta\bar{\zeta}\right)u_0,$$
$$u_{\beta+1} = 2\,\mathrm{Re}\left(z^2\bar{\zeta}\right)u_\beta - \left(z^2\bar{z}^2\zeta\bar{\zeta}\right)u_{\beta-1}, \qquad (\beta \geqq 1),$$
$$u_1 = 2\,\mathrm{Re}\left(z^2\bar{\zeta}\right)u_0 - \left(z^2\bar{z}^2\right)x_1.$$

Hence inductively the only new generators are t_1 and u_0; that is,

$$\begin{aligned} &\bar{\zeta}H_k+\zeta\overline{H}_k, \qquad k=3,4,\\ &\bar{z}^2H_k+z^2\overline{H}_k, \qquad k=3,4. \end{aligned} \tag{A.44}$$

However, we observe that the identities

$$\begin{aligned} N\,\mathrm{Re}(H_1)-2\delta\,\mathrm{Im}(H_2)&=\mathrm{Re}(\bar{z}H_3),\\ 2\delta\,\mathrm{Im}(H_1)+N\,\mathrm{Re}(H_2)&=\mathrm{Re}(\bar{z}H_4) \end{aligned}$$

are valid. Thus the generators $\bar{z}^2H_k+z^2\overline{H}_k$ $(k=3,4)$ are redundant and can be omitted.

From (A.34) we have either $\alpha=0$, $\beta=1$ or $\alpha=2$, $\beta=0$. These lead to t_1 and u_0 again.

For convenience we now consider (A.38), for which $\alpha=2$, $\beta=0$ or $\alpha=0$, $\beta=1$. These lead to new generators

$$\begin{aligned} &\delta\,\mathrm{Im}(\bar{\zeta}H_k), \qquad k=3,4,\\ &\delta\,\mathrm{Im}(\bar{z}^2H_k), \qquad k=3,4. \end{aligned} \tag{A.45}$$

Finally we take (A.36) and (A.37), yielding $\alpha=2\beta-2$ $(\beta\geqq 1)$ and $\alpha=2\beta+2$ respectively. So we have terms

$$\begin{aligned} r_\beta&=i\delta z^{2\beta-2}\bar{\zeta}^\beta H_k-i\delta\bar{z}^{2\beta-2}\zeta^\beta\overline{H}_k \qquad (\beta\geqq 1),\\ s_\beta&=i\delta\bar{z}^{2\beta-2}\zeta^\beta H_k-i\delta z^{2\beta-2}\bar{\zeta}^\beta\overline{H}_k \qquad (\beta\geqq 0). \end{aligned}$$

As usual, we find that

$$\begin{aligned} r_{\beta+1}&=2\,\mathrm{Re}(z^2\bar{\zeta})r_\beta-(z^2\bar{z}^2\zeta\bar{\zeta})r_{\beta-1} \qquad (\beta\geqq 2),\\ r_2&=2\,\mathrm{Re}(z^2\bar{\zeta})\delta\,\mathrm{Im}(\bar{\zeta}H_k)-(\zeta\bar{\zeta})\delta\,\mathrm{Im}(\bar{z}^2H_k),\\ s_{\beta+1}&=2\,\mathrm{Re}(z^2\bar{\zeta})s_\beta-(z^2\bar{z}^2\zeta\bar{\zeta})s_{\beta-1},\\ s_1&=2\,\mathrm{Re}(z^2\bar{\zeta})s_0-(z^2\bar{z}^2)\delta\,\mathrm{Im}(\bar{\zeta}H_k),\\ s_0&=\delta\,\mathrm{Im}(\bar{z}^2H_k). \end{aligned}$$

Taking (A.45) into account, we find no new generators.

This completes the analysis. We have found twelve generators (A.39), (A.42), (A.43), (A.44), (A.45). Proposition A.1(b) now follows.

Note that the invariants (A.2) for $O(2)\times S^1$ do not form a polynomial ring: there is a relation

$$\delta^2\gamma^2-\sigma^2=(z\bar{z})^2(\zeta\bar{\zeta})=\beta^2(N-4\delta^2).$$

Further, the equivariants do not form a free module, although the relations have degree 9 or more. For example

$$[\sigma][\delta\,\mathrm{Im}(H_1)]=[\delta^2][\mathrm{Im}(z^2\bar{\zeta})\mathrm{Im}(H_1)].$$

(There are other relations too). In consequence, the singularity theory of $O(2)\times S^1$ on the six-dimensional kernel would be extremely complicated to compute.

Finally, we turn to the statement of Theorem 4.3. We obtain the form stated there for the equivariants from that used in Proposition A.1, by defining

$$K_j = \mathrm{Re}(H_j), \quad L_j = \mathrm{Im}(H_j), \quad j = 1,2,3,4,$$

and collecting terms according to the matrices K_j, L_j that occur.

Acknowledgments. We are grateful to Bill Langford for suggesting that group theory might be useful in the analysis of the six-dimensional kernel. We thank Mike Gorman for helping to interpret the experimental evidence and David Andereck for sharing his unpublished observations. John Guckenheimer pointed out some redundancies in our original list of generators for the module of equivariant mappings, thus simplifying a number of calculations.

This work was carried out while the second author held a visiting position in the Mathematics Department, University of Houston.

REFERENCES

C. D. ANDERECK [1984], *Private communication.*

C. D. ANDERECK, R. D. DICKMAN AND H. L. SWINNEY [1983], *New flows in a circular Couette system with co-rotating cylinders*, Phys. Fluids, 26, pp. 1395–1401.

C. D. ANDERECK, S. S. LIU AND H. L. SWINNEY [1984], *Flow regimes in a circular Couette system with independently rotating cylinders*, preprint.

T. B. BENJAMIN [1978a], *Bifurcation phenomena in steady flow of a viscous fluid. I. Theory*, Proc. Roy. Soc. London A, 359, pp. 1–26.

_____ [1978b], *Bifurcation phenomena in a steady flow of a viscous fluid. II. Experiments*, Proc. Roy. Soc. London A, 359, pp. 27–43.

T. B. BENJAMIN AND T. MULLIN [1982], *Notes on the multiplicity of flows in the Taylor experiment*, J. Fluid Mech. 121, pp. 219–230.

P. CHOSSAT [1985], *Interaction d'ondes rotatires dans le problème de Couette–Taylor*, C. R. Acad. Sci. Paris, 300, Ser. I, no. 8, pp. 251–254.

P. CHOSSAT AND G. IOOSS [1984], *Primary and second bifurcation in the Couette–Taylor problem*, Université de Nice, Preprint.

D. COLES [1965], *Transition in circular Couette flow*, J. Fluid Mech., 93, pp. 515–527.

A. DAVEY [1962], *The growth of Taylor vortices in flow between rotating cylinders*, J. Fluid Mech. 14, pp. 336–368.

A. DAVEY, R. C. DIPRIMA AND J. T. STUART [1968], *On the instability of Taylor vortices*, J. Fluid Mech., 31, pp. 17–52.

R. C. DIPRIMA, P. M. EAGLES AND J. SIJBRAND [1984], *Interaction of axisymmetric and nonaxisymmetric disturbances in the flow between concentric counter-rotating cylinders*, in preparation.

R. C. DIPRIMA AND R. N. GRANNICK [1971], *A nonlinear investigation of the stability of flow between counter-rotating cylinders*, in *Instability of Continuous Systems*, H. Leipholz, ed., Springer-Verlag, Berlin, pp. 55–60.

R. C. DIPRIMA AND J. SIJBRAND [1982], *Interactions of axisymmetric and non-axisymmetric disturbances in the flow between concentric rotating cylinders: Bifurcations near multiple eigenvalues*, in Stability in the Mechanics of Continua, F. H. Schroeder, ed., Springer-Verlag, Berlin, pp. 383–386.

R. C. DIPRIMA AND H. L. SWINNEY [1981], *Instabilities and transition in flow between concentric rotating cylinders*, in *Hydrodynamic Instabilities and the Transition to Turbulence*, H. L. Swinney and J. P. Gollub, eds., Topics in Applied Physics 45, Springer-Verlag, Berlin, pp. 139–180.

J. P. GOLLUB AND H. L. SWINNEY [1975], *Onset of turbulence in a rotating fluid*, Phys. Rev. Lett., 35, pp. 927–930.

M. GOLUBITSKY AND I. N. STEWART [1985], *Hopf bifurcation in the presence of symmetry*, Arch. Rational Mech. Anal., 87, pp. 107–165.

M. GORMAN, L. A. REITH AND H. L. SWINNEY [1980], *Modulation patterns, multiple frequencies, and other phenomena in circular Couette flow*, in Nonlinear Dynamics, R. Helleman, ed., Ann. N. Y. Acad. Sci., 357, pp. 10–21.

M. GORMAN, H. L. SWINNEY AND D. A. RAND [1981], *Doubly periodic circular Couette flow: Experiments compared with predictions from dynamics and symmetry*, Phys. Rev. Lett., 46, pp. 992–995.

E. R. KRUEGER, A. GROSS AND R. C. DIPRIMA [1966], *On the relative importance of Taylor-vortex and nonaxisymmetric modes in flow between rotating cylinders*, J. Fluid Mech. 24, pp. 521–538.

J. SCHEURLE AND J. E. MARSDEN [1984], *Bifurcation to quasi-periodic tori in the interaction of steady state and Hopf bifurcations*, this Journal, pp. 1055–1074.

V. POÉNARU, *Singularités C^∞ en présence de symétrie*, Lecture Notes in Mathematics 510, Springer-Verlag, Berlin, 1976.

D. RAND [1982], *Dynamics and symmetry: predictions for modulated waves in rotating fluids*, Arch. Rational Mech. Anal., 79, pp. 1–38.

D. H. SATTINGER [1983], *Branching in the Presence of Symmetry*, CBMS Regional Conference Series in Applied Mathematics, Society for Industrial and Applied Mathematics, Philadelphia.

S. SCHECTER [1976], *Bifurcations with symmetry*, in The Hopf Bifurcation and its Applications, J. E. Marsden and M. McCracken, eds., Appl. Math. Sci., 19, Springer-Verlag, New York, pp. 224–249.

G. SCHWARZ [1975], *Smooth functions invariant under the action of a compact group*, Topology, 14, pp. 63–68.

R. S. SHAW, C. D. ANDERECK, L. A. REITH, AND H. L. SWINNEY [1982], *Superposition of travelling waves in the circular Couette system*, Phys. Rev. Lett., 48, pp. 1172–1175.

G. I. TAYLOR [1923], *Stability of a viscous liquid contained between two rotating cylinders*, Phil. Trans. Roy. Soc. London A, 223, pp. 289–343.

SIAM J. MATH. ANAL.
Vol. 17, No. 2, March, 1986

002

BIFURCATIONS IN DOUBLY-DIFFUSIVE SYSTEMS III. INTERACTION OF EQUILIBRIUM AND TIME PERIODIC SOLUTIONS*

WAYNE NAGATA† AND JAMES W. THOMAS‡

Abstract. The bifurcations associated with a double zero eigenvalue which occur in equations describing double-diffusive convection in a layer of fluid are studied. Two-dimensional roll-like, and three-dimensional rectangular, square and hexagonal convection cell patterns are considered. The equations are reduced to a center manifold and further reduced to a normal form so that the complete unfolding of the codimension two bifurcation can be determined. In addition to primary pitchfork and Hopf bifurcations, saddle connections exist for roll-like, square and hexagonal convective solutions. For certain parameter values, secondary Hopf bifurcations and semistable periodic orbits exist for square and hexagonal convective solutions.

Key words. double-diffusive convection, codimension two bifurcation

1. Introduction. This is the final paper in a series of three concerning bifurcations in double-diffusive convective equations. The equations model an idealized infinite horizontal layer of fluid that is heated and salted from below. For more background we refer to the first paper of this series [8], which we will call Part I, and the references therein. In Part I we treated the bifurcation of equilibrium solutions corresponding to steady convective cells, and in the second paper of the series [9], which we will call Part II, we treated the Hopf bifurcation of time periodic solutions corresponding to oscillating (overstable) convection cells. In the present paper we consider the interactions between equilibrium and time periodic solutions and show the existence of further bifurcations such as saddle connections and secondary Hopf bifurcations, in some cases (see Figs. 3.4 and 3.6).

We restrict the convection equations to classes of functions corresponding to prescribed cellular convection patterns. The linearization of the restricted equations can then have a nilpotent double zero eigenvalue, and the associated codimension two bifurcation is analyzed using center manifold and normal form reductions to obtain our results. However, our stability results are incomplete since we consider only local asymptotic stability within a prescribed class of convection cell pattterns. For example, our results may show that hexagon pattern solutions are locally asymptotically stable with respect to hexagon pattern disturbances, but more generally they may well be unstable with respect to roll-like disturbances. By considering instead a wider class of solutions, namely those which are doubly periodic with respect to a hexagonal lattice, one can treat rolls, hexagons, triangles and a class of rectangles simultaneously to obtain pattern selection results. Motivated by the Rayleigh–Bénard problem modelling an infinite horizontal layer of fluid heated from below (no salt), Golubitsky, Swift and Knobloch [15] developed a theory of pattern selection in the hexagonal lattice. Bifurcations of equilibrium solutions associated with a six-dimensional semisimple eigenspace were treated. The third-order coefficients computed in Part I were sufficient (assuming a generically satisfied condition involving fifth-order coefficients) to determine some pattern selection results in the hexagonal lattice for steady convection. To treat pattern

* Received by the editors February 10, 1984, and in revised form November 26, 1984.
† Present address: Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912.
‡ Department of Mathematics, Colorado State University, Fort Collins, Colorado 80513.

selection in the hexagonal lattice in the context of the present paper, one must analyze the dynamic bifurcations associated with a twelve-dimensional nilpotent eigenspace. This is a substantial project and we do not attempt it here. Should such a project be completed, however, the third-order normal form coefficients computed in this paper would be expected to be incorporated into the hexagonal lattice normal form to give specific pattern selection results as in Part I.

In any case, results for the hexagonal lattice do not include square and most rectangular convection cell patterns. Sattinger [16] includes a discussion of square and rectangular lattices, but so far there are no results comparing the stability of solutions in one lattice with respect to disturbances in another. Furthermore, we do not consider stability of solutions with respect to more general disturbances, for example spatially periodic disturbances with different wave numbers. For the stability of rolls in the Rayleigh–Bénard problem with respect to a wide variety of disturbances, see Busse [13]. Finally, we do not consider the effects of sidewalls in a finite layer of fluid. These are perhaps the most important effects (and most difficult to treat mathematically) which must be considered in any complete treatment of pattern selection. For some results on rolls, in the Raleigh–Bénard problem, see Daniels [14].

The plan of this paper is as follows: First, in this section we review the parametrized system of partial differential equations modelling double-diffusive convection, and the symmetry conditions corresponding to cellular convection patterns. We then consider the eigenvalue problem for the linearization, linearized about the constant gradient solution. For certain critical parameter values, the linearization has a double zero eigenvalue when restricted to functions satisfying the symmetry conditions. We then reformulate the problem in terms of operators in Hilbert spaces. The operators and spaces are equivalent to those used in Parts I and II, but instead are formulated in terms of Fourier series. In §2 we apply a version of the center manifold theorem due to Henry [5] to reduce the parametrized partial differential equations to a two-parameter family of ordinary differential equations on two-dimensional invariant manifolds. All of the dynamics in the original family of partial differential equations for parameter values near the critical ones are retained by the family of reduced ordinary differential equations. We then further reduce the equations to a normal form, from which the complete local bifurcation and stability behavior of solutions to the equations can be deduced. In §3 we present the results of our computations to find the normal form coefficients. These coefficients determine the unfolding of the degenerate vector field in the neighborhood of critical parameter values. Two-dimensional roll-like solutions can exhibit only one case of the unfolding, but we find that three-dimensional cellular solutions (rectangular, square or hexagonal) can exhibit further cases depending on the values of auxiliary parameters.

We recall from Parts I and II that the deviation $u=(u_1,u_2,u_3,u_4,u_5)$ from the constant gradient solution satisfies the system of coupled partial differential equations

$$
\begin{aligned}
&\frac{\partial}{\partial t}\mathbf{u}=\sigma(\Delta\mathbf{u}-\nabla p)+(r\sigma u_4-s\sigma u_5)\mathbf{e}_3-(\mathbf{u}\cdot\nabla)\mathbf{u},\\
&\frac{\partial}{\partial t}u_4=\Delta u_4+u_3-(\mathbf{u}\cdot\nabla)u_4,\\
&\frac{\partial}{\partial t}u_5=\tau\Delta u_5+u_3-(\mathbf{u}\cdot\nabla)u_5,\\
&\operatorname{div}\mathbf{u}=0,
\end{aligned}
\tag{1.1}
$$

for $\mathbf{x}=(x_1,x_2,x_3)$ in $\mathbb{R}^2\times(0,1)$, where $\mathbf{e}_3$ is the unit vector $(0,0,1)$, $\mathbf{u}=(u_1,u_2,u_3)$ is the nondimensionalized fluid velocity, u_4 is the nondimensionalized fluid temperature deviation from the constant gradient solution and u_5 is the nondimensionalized solute concentration deviation from the constant gradient solution. The parameters r, s, σ and τ are all positive, with $0<\tau<1$. We take free-surface boundary conditions for $\mathbf{u}$ and Dirichlet boundary conditions for u_4 and u_5 at the boundary surfaces $x_3=0$ and $x_3=1$:

$$\left.\frac{\partial u_1}{\partial x_3}\right|_{x_3=0,1}=\left.\frac{\partial u_2}{\partial x_3}\right|_{x_3=0,1}=u_3|_{x_3=0,1}=u_4|_{x_3=0,1}=u_5|_{x_3=0,1}=0. \tag{1.2}$$

We require that solutions of (1.1)–(1.2) are periodic with fundamental domain of spatial periodicity Ω, where Ω corresponds to roll-like, rectangular, square or hexagonal convection cells, as described below. Then we can express the functions u_k, $k=1,\cdots,5$ and p as Fourier series (cf. Part I)

$$w(\mathbf{x})=\sum_{\mathbf{j}\in\mathbb{Z}^n}\hat{w}(\mathbf{j})e^{i(j_1\alpha_1x_1+\cdots+j_3\alpha_3x_3)},\qquad \bar{\hat{w}}(\mathbf{j})=\hat{w}(-\mathbf{j}), \tag{1.3}$$

where $\mathbf{j}=(j_1,j_2,j_3)$, $w=u_k$, $k=1,\cdots,5$ or p, $n=2$ or 3, $\alpha_1^2+\alpha_2^2=\pi^2/2$ and $\alpha_3=\pi$. The Fourier coefficients u_k, p are required to satisfy symmetry conditions

$$\begin{aligned}
&\hat{u}_1(\mathbf{j}) \text{ is odd in } j_1, \quad \text{even in } j_2, \quad \text{even in } j_3,\\
&\hat{u}_2(\mathbf{j}) \text{ is even in } j_1, \quad \text{odd in } j_2, \quad \text{even in } j_3,\\
&\hat{u}_3(\mathbf{j}) \text{ is even in } j_1, \quad \text{even in } j_2, \quad \text{odd in } j_3,\\
&\hat{w}(\mathbf{j}) \text{ is even in } j_1, \quad \text{even in } j_2, \quad \text{odd in } j_3,
\end{aligned} \tag{1.4}$$

where $\hat{w}=\hat{u}_4$, $\hat{u}_5$ or $\hat{p}$, and

$$\hat{\mathbf{u}}(g\mathbf{j})=g\hat{\mathbf{u}}(\mathbf{j}),\qquad \hat{w}(g\mathbf{j})=\hat{w}(\mathbf{j}), \tag{1.5}$$

where $\hat{w}=\hat{u}_4$, $\hat{u}_5$ or $\hat{p}$, and g is a 3×3 matrix of the form

$$g=\begin{bmatrix}\cos\phi & -\sin\phi & 0\\ \sin\phi & \cos\phi & 0\\ 0 & 0 & 1\end{bmatrix}$$

representing rotations by the angle ϕ about the x_3 axis.

We define the following classes of functions corresponding to the cellular convection patterns (cf. Part I):

(a) *Rolls.* $\alpha_1=\pi/2^{1/2}$, $\alpha_2=0$ in (1.3). For rolls only, we modify the definitions of $\mathbf{x}$, $\mathbf{u}$ and $\mathbf{j}$ to $\mathbf{x}=(x_1,x_3)$, $\mathbf{u}=(u_1,u_3)$ and $\mathbf{j}=(j_1,j_3)$. The sum in (1.3) is over $\mathbf{j}\in\mathbb{Z}^2$ for rolls ($n=2$). $\phi=\pi$ in (1.5) and $\Omega=(0,2\pi/\alpha_1)\times(0,1)$.

(b) *Rectangles.* $\alpha_1^2+\alpha_2^2=\pi^2/2$, $\alpha_1,\alpha_2>0$, $\alpha_2\neq\alpha_1[m^2-1]^{1/2}$ for any positive integer m in (1.3). $\phi=\pi$ in (1.5) and $\Omega=(0,2\pi/\alpha_1)\times(0,2\pi/\alpha_2)\times(0,1)$.

(c) *Squares.* $\alpha_1=\alpha_2=\pi/2$ in (1.3). $\phi=\pi/2$ in (1.5) and $\Omega=(0,2\pi/\alpha_1)\times(0,2\pi/\alpha_1)\times(0,1)$.

(d) *Hexagons.* $\alpha_1=2^{1/2}\pi/4$, $\alpha_2=6^{1/2}\pi/4$ in (1.3). $\phi=2\pi/3$ in (1.5) and $\Omega=H\times(0,1)$, where H is the open hexagonal region in $\mathbb{R}^2$ enclosed by the six lines $x_2=\pm\pi/3^{1/2}\alpha_1$, $x_2+3^{1/2}x_1=\pm2\pi/3^{1/2}\alpha_1$, $x_2-3^{1/2}x_1=\pm2\pi/3^{1/2}\alpha_1$.

We observe that by (1.4) we have

$$\hat{u}_k(\mathbf{0})=0,\quad k=1,\cdots,5,\qquad \hat{p}(\mathbf{0})=0 \tag{1.6}$$

and furthermore, all the $\hat{u}_k(\mathbf{j})$, $\hat{p}(\mathbf{j})$ are pure imaginary. Thus, if desired we can write the Fourier series (1.3) as (cf. Part I)

$$u_k(\mathbf{x}) = \sum_{j_1,j_2=-\infty}^{\infty} \left[\hat{u}_k(j_1,j_2,0) + \sum_{j_3=1}^{\infty} 2\hat{u}_k(\mathbf{j}) \cos j_3 \pi x_3 \right] e^{i(j_1\alpha_1 x_1 + j_2\alpha_2 x_2)}$$

for $k=1,2$ and

$$(\mathbf{x}) = \sum_{j_1,j_2=-\infty}^{\infty} \left[\sum_{j_3=1}^{\infty} 2i\hat{w}(\mathbf{j}) \sin j_3 \pi x_3 \right] e^{i(j_1\alpha_1 x_1 + j_2\alpha_2 x_2)}$$

where $w=u_3$, u_4, u_5 or p. For rolls there are slight modifications ($\alpha_2=0$, $\mathbf{u}=(u_1,u_3)$ sum only over j_1,j_3).

If we substitute the Fourier series (1.3) with the Fourier coefficients satisfying the symmetries corresponding to one of the cellular convection patterns (a)–(d) into the eigenvalue problem for the linearization

$$\begin{aligned} &\sigma(\Delta\mathbf{u} - \nabla p) + (r\sigma u_4 - s\sigma u_5)\mathbf{e}_3 = \lambda\mathbf{u}, \\ &\Delta u_4 + u_3 = \lambda u_4, \\ &\tau\Delta u_5 + u_3 = \lambda u_5, \\ &\operatorname{div}\mathbf{u} = 0, \end{aligned} \tag{1.7}$$

with the boundary conditions (1.2), we find that the critical eigenvalues λ belong to the lowest mode and are given by the roots of the cubic equation

$$\begin{aligned} \lambda^3 + (3\pi^2/2)(1+\sigma+\tau)\lambda^2 + \left[(9\pi^4/4)(\sigma+\tau+\sigma\tau) - (1/3)\sigma(r-s)\right]\lambda \\ + (27\pi^6/8)\sigma\tau + (\pi^2/2)\sigma(s-\tau r) = 0 \end{aligned} \tag{1.8}$$

with the same multiplicities. Higher modes correspond to eigenvalues whose real parts are negative when (1.8) has solutions λ with real part near 0. For $\sigma>0$, $0<\tau<1$ and for the critical parameter values

$$\begin{aligned} r = r_0 &\equiv (27\pi^4/4)\sigma^{-1}(1-\tau)^{-1}(\sigma+\tau), \\ s = s_0 &\equiv (27\pi^4/4)\tau^2(1-\tau)^{-1}(1+\sigma^{-1}) \end{aligned} \tag{1.9}$$

the equation (1.8) has a double root $\lambda=0$ and the remaining root is real and negative (cf. Part I).

Thus for $\sigma>0$, $0<\tau<1$ and for each Ω corresponding to roll-like, rectangular, square or hexagonal convection cells, the eigenvalue problem (1.7) for u_k, $k=1,\cdots,5$ and p satisfying (1.3)–(1.5) has a double zero eigenvalue when (1.9) holds. All other eigenvalues have negative real parts for these critical parameter values.

We now reformulate the eigenvalue problem (1.7) in terms of an abstract differential operator in a space of functions. Let $n=2$ (for rolls) or $n=3$ (for rectangles, squares or hexagons) and define the spaces of functions corresponding to one of the cellular convection patterns (a)–(d)

$$\dot{H}^{\beta}_{\#}(\Omega) = \left\{ w : w(\mathbf{x}) \text{ has the form (1.3)}, \sum_{\mathbf{j}\in\mathbb{Z}^n} |\mathbf{j}|^{2\beta} |\hat{w}(\mathbf{j})|^2 < \infty, \hat{w}(\mathbf{j}) \text{ satisfy (1.4) and (1.5)} \right\}$$

and

$$\dot{\mathbf{H}}^{\beta}_{\#}(\Omega)=\Big\{\mathbf{u}: u_k(\mathbf{x}) \text{ have the form (1.3)}, k=1,\cdots,3,$$

$$\sum_{\mathbf{j}\in\mathbb{Z}^n}|\mathbf{j}|^{2\beta}|\hat{\mathbf{u}}(\mathbf{j})|^2<\infty, \hat{\mathbf{u}}(\mathbf{j}) \text{ satisfy (1.4) and (1.5)}\Big\}$$

where $\beta\in\mathbb{R}$, and the numbers α_1, α_2, α_3 in (1.3) and the matrix g in (1.5) are appropriate to rolls, rectangles, squares or hexagons. The function spaces $\dot{H}^{\beta}_{\#}(\Omega)$ and $\dot{\mathbf{H}}^{\beta}_{\#}(\Omega)$ are closed subspaces of the Hilbert spaces $\dot{H}^{\beta}_{p}(Q)$ and $\dot{\mathbb{H}}^{\beta}_{p}(Q)$ defined in [12] with $Q=(0,2\pi/\alpha_1)\times(0,2\pi/\alpha_3)$ for rolls or $Q=(0,2\pi/\alpha_1)\times(0,2\pi/\alpha_2)\times(0,2\pi/\alpha_3)$ for rectangles, squares, or hexagons; hence the statements made in [12, pp. 7–11] regarding $\dot{H}^{\beta}_{p}(Q)$, $\dot{\mathbb{H}}^{\beta}_{p}(Q)$ and the Stokes operator A carry over with minor modifications to our present case.

In particular, $\dot{H}^{\beta}_{\#}(\Omega)$ is a Hilbert space with norm

$$\|w\|_{\beta}=\left[\sum_{\mathbf{j}\in\mathbb{Z}^n}|\mathbf{j}|^{2\beta}|\hat{w}(\mathbf{j})|^2\right]^{1/2}$$

and $\dot{\mathbf{H}}^{\beta}_{\#}(\Omega)$ is a Hilbert space with the product structure of $[\dot{H}^{\beta}_{\#}(\Omega)]^n$. The closed subspace

$$J=\{\mathbf{u}\in\dot{\mathbf{H}}^{0}_{\#}(\Omega):\operatorname{div}\mathbf{u}=0\}$$

of $\dot{H}^{0}_{\#}(\Omega)$ has the orthogonal complement

$$G=\{\mathbf{u}\in\dot{\mathbf{H}}^{0}_{\#}(\Omega):\mathbf{u}=\nabla p \text{ for some } p\in\dot{H}^{1}_{\#}(\Omega)\}.$$

Let $X=J\times\dot{H}^{0}_{\#}(\Omega)\times\dot{H}^{0}_{\#}(\Omega)$ with the norm induced by the product structure of $[\dot{H}^{0}_{\#}(\Omega)]^{n+2}$. By solving the boundary-value problem

$$\begin{aligned}&\sigma(\Delta\mathbf{u}-\nabla p)=\mathbf{f},\\&\Delta u_4=f_4,\\&\tau\Delta u_5=f_5,\\&\operatorname{div}\mathbf{u}=0\end{aligned}\tag{1.10}$$

with boundary values (1.2) explicitly for the Fourier coefficients of $u=(\mathbf{u},u_4,u_5)$ when $\sigma>0$, $0<\tau<1$ and $f=(\mathbf{f},f_4,f_5)\in X$ as in [12], it follows that $p=0$ and that the operator defined by

$$\begin{aligned}&D(A)=X\cap\left[\dot{H}^{2}_{\#}(\Omega)\right]^{n+2},\\&Au=(\sigma\Delta\mathbf{u},\Delta u_4,\tau\Delta u_5)\quad \text{for } u\in D(A)\end{aligned}\tag{1.11}$$

is a closed, densely defined selfadjoint operator in X with compact inverse $A^{-1}:X\to X$. The norm

$$|u|_1=\|Au\|_X$$

on $D(A)$ is equivalent to the norm induced by the product structure of $[\dot{H}^2_{\#}(\Omega)]^{n+2}$. The operator $-A$ is sectorial (cf. [5, p. 19]), and hence we can define the powers $(-A)^\alpha$ for all real $\alpha \geqq 0$ and the Hilbert spaces

$$X^\alpha = D\big((-A)^\alpha\big)$$

with norms

$$|u|_\alpha = \|(-A)^\alpha u\|_X.$$

We have $D(A) = X^1 \subset X^\alpha \subset X^0 = X$ for $0 \leqq \alpha \leqq 1$, the inclusions being continuous. In terms of the spaces $\dot{H}^\beta_{\#}(\Omega)$ we have

$$X^\alpha = \left\{ u = (\mathbf{u}, u_4, u_5) \in \dot{\mathbf{H}}^{2\alpha}_{\#}(\Omega) \times \dot{H}^{2\alpha}_{\#}(\Omega) \times \dot{H}^{2\alpha}_{\#}(\Omega) : \operatorname{div} \mathbf{u} = 0 \right\}$$

and the norm $|u|_\alpha$ is equivalent to the norm by the product structure of $[\dot{H}^{2\alpha}_{\#}(\Omega)]^{n+2}$.

Let Π denote the orthogonal projection of $\dot{\mathbf{H}}^0_{\#}(\Omega)$ onto J, and let

$$\Pi_0 = \Pi \times I \times I : J \times \dot{H}^0_{\#}(\Omega) \times \dot{H}^0_{\#}(\Omega) \to X,$$

where I is the identity mapping on $\dot{H}^0_{\#}(\Omega)$. Then Π_0 is a continuous linear mapping, and for $\sigma > 0$, the operator

$$B(r,s)u = \Pi_0\big((r\sigma u_4 - s\sigma u_5)\mathbf{e}_3, u_3, u_3\big) \tag{1.12}$$

is a continuous linear mapping in u, from X^α into X, for each fixed pair $(r,s) \in \mathbb{R}^2$. Furthermore, $B(r,s)$ depends analytically on (r,s).

We let

$$L(r,s)u = Au + B(r,s)u, \qquad u \in D(L(r,s)) = D(A). \tag{1.13}$$

Since A has a compact inverse, the argument in the Appendix of Part I can be used to show that the spectrum of $L(r,s)$ consists entirely of isolated eigenvalues of finite multiplicities. The eigenvalue problem (1.7) can now be written as

$$L(r,s)u = \lambda u, \qquad u \in D(L(r,s)). \tag{1.14}$$

The eigenvalues of (1.14) corresponding to the lowest mode are given by (1.8). Thus $L(r_0, s_0)$ has a double zero eigenvalue, and the rest of the spectrum of $L(r_0, s_0)$ consists of isolated eigenvalues with negative real parts. Note that the space X corresponds to a particular choice of cellular convection—rolls, rectangles, squares or hexagons.

Finally, we express the nonlinear double-diffusive convection equations in abstract form. From the argument given in [5, pp. 79–81] and from the continuity of Π_0, it follows that the operator M: $X^\alpha \times X^\alpha \to X$, defined by

$$M(u,v) = -\Pi_0\big((\mathbf{u}\cdot\nabla)\mathbf{v}, (\mathbf{u}\cdot\nabla)v_4, (\mathbf{u}\cdot\nabla)v_5\big) \tag{1.15}$$

is a continuous bilinear mapping for $3/4 < \alpha < 1$. Thus

$$f(u,r,s) = B(r,s)u + M(u,u) \tag{1.16}$$

is analytic mapping f: $X^\alpha \times \mathbb{R}^2 \to X$. We write the double-diffusive convection equations (1.1), with the boundary conditions (1.2) and one of the cellular convection patterns (a)–(d) as

$$\frac{du}{dt} = Au + f(u,r,s) \tag{1.17}$$

where $-A$ is sectorial and f is analytic. Furthermore, the linear part of the right-hand side of (1.17) is

$$Au + D_u f(0,r,s)u = L(r,s)u, \tag{1.18}$$

where $D_u f$ is the partial derivative of f with respect to u.

We note that (1.17) generates a unique parametrized family of local semiflows near the origin in X^α, $\frac{3}{4} < \alpha < 1$, which are the solutions of (1.17) with initial conditions $u(0) = u_0 \in X^\alpha$ [5, p. 54]. This family of semiflows is jointly analytic in (t, u_0, r, s) for $t > 0$ on its domain of existence [5, p. 66].

2. Center manifold reduction. The study of bifurcation and stability in differential equations can often be greatly simplified by the use of the center manifold theorem. This theorem allows one to reduce the dimension of the state space while preserving the local behavior of solutions of a differential equation. As an example, consider the ordinary differential equation

$$\frac{du}{dt} = Lu + g(u), \qquad u \in \mathbb{R}^m, \tag{2.1}$$

where L is an $m \times m$ matrix, g: $\mathbb{R}^m \to \mathbb{R}^m$ is a smooth nonlinear mapping with $g(0) = 0$ and $g'(0) = 0$. Then the origin $u = 0$ is an equivalent solution of (2.1), and to study the stability of the origin we determine the spectrum of L, $\Sigma(L)$. Suppose that the origin is a degenerate equilibrium of (2.1), i.e. $\Sigma(L)$ contains eigenvalues with zero real part. For example, if $\Sigma(L)$ consists of m_c eigenvalues with zero real part and m_s eigenvalues with negative real part, $m_c + m_s = m$, then the phase space $\mathbb{R}^m$ splits into a direct sum of L-invariant subspaces Y_c (the center eigenspace) and Y_s (the stable eigenspace)

$$\mathbb{R}^m = Y_c \oplus Y_s$$

where $\dim Y_c = m_c$ and $\dim Y_s = m_s$, and thus every $u \in \mathbb{R}^m$ can be uniquely expressed as

$$u = u_c + u_s, \quad \text{for some } u_c \in Y_c, \quad u_s \in Y_s.$$

Furthermore, the restrictions $L \mid_{Y_c}$ and $L \mid_{Y_s}$ have spectra consisting of the eigenvalues of L with zero real parts, and the eigenvalues of L with negative real parts, respectively. Provided g is smooth enough, the center manifold theorem then states that there is a (not necessarily unique) local invariant manifold W_c—a submanifold of $\mathbb{R}^m$ defined in some neighborhood of the origin, consisting of solution curves of (2.1)—tangent to Y_c at the origin

$$W_c = \{ u = u_c + u_s \in \mathbb{R}^m : u_s = h(u_c), |u_c| < \delta \}$$

where h is a smooth function from a neighborhood of the origin in Y_c into Y_s, with $h(0) = 0$ and $h'(0) = 0$. W_c is called a center manifold.

The asymptotic behavior of solutions of (2.1) near the origin is determined by the m_c-dimensional equation in Y_c

$$\frac{du_c}{dt} = L_c u_c + P_c g(u_c + h(u_c)), \tag{2.2}$$

where $L_c = L \mid_{Y_c}$ and P_c is the projection of $\mathbb{R}^m$ onto Y_c, along Y_s. More precisely, if the origin is stable (asymptotically stable, unstable) for (2.2), then the origin is stable (asymptotically stable, unstable) for (2.1) [1, p. 4]. Thus, center manifold theory allows one to study the stability of an equivalent solution by locally reducing the differential equation to one of lower dimension.

To study bifurcations near the origin in a k-parameter family of differential equations

$$\frac{du}{dt} = L(\mu)u + g(u,\mu), \qquad u \in \mathbb{R}^m, \quad u \in \mathbb{R}^k, \tag{2.3}$$

where $L(\mu)$ is an $m \times m$ matrix depending smoothly on μ, $L(0)$ has m_c eigenvalues with zero real part and m_s eigenvalues with negative real part, $m_c + m_s = m$, $g(0,\mu) = 0$ and $D_u g(0,\mu) = 0$, we apply the center manifold to the $(m+k)$-dimensional system

$$\frac{du}{dt} = L(\mu)u + g(u,\mu), \qquad \frac{d\mu}{dt} = 0 \tag{2.4}$$

for $(u,\mu) \in \mathbb{R}^{m+k}$. $\mathbb{R}^m$ splits into a direct sum of $L(0)$-invariant subspaces $\mathbb{R}^m = X_c \oplus X_s$, where $\mathrm{Re}\,\Sigma(L(0) \mid_{X_c}) = 0$, $\mathrm{Re}\,\Sigma(L(0) \mid_{X_s}) < 0$; and $\mathbb{R}^{m+k}$ splits into a direct sum $\mathbb{R}^{m+k} = Y_c \oplus Y_s$, where $Y_c = X_c \times \mathbb{R}^k$, $Y_s = X_s$. The center manifold theorem applied to (2.4) gives the existence of a center manifold defined by a smooth function $h(u_c,\mu)$ mapping a neighborhood of the origin in $X_c \times \mathbb{R}^k$ into X_s. Bifurcation and stability of solutions of (2.3) near the origin, for μ near 0, is determined by the reduced system

$$\frac{du_c}{dt} = P_c L(\mu)u_c + P_c g(u_c + h(u_c,\mu),\mu) \tag{2.5}$$

for $u_c \in X_c$, $\mu \in \mathbb{R}^k$ [1, p. 12], [6, pp. 471–473]. Thus local bifurcations in (2.3) can be studied by means of a reduced system (2.5). Moreover, in many applications it suffices to determine only the first few terms in the Taylor series expansion of the right-hand side of (2.5) in order to obtain the complete unfolding of (2.3) near $\mu = 0$.

In this section we apply a suitable version of the center manifold theorem to reduce the (r,s)-parametrized family of abstract differential equations (1.17) to a two-parameter family of ordinary differential equations in a two-dimensional phase space, the dimension two of the phase space being determined by the multiplicity two of the zero eigenvalue of the linearization (1.14). We explicitly describe X_c, P_c and the first few terms in the Taylor series expansion of a center manifold function h when the state space X of (1.17) corresponds to rolls, rectangles, squares or hexagons.

We apply the following theorems due to Henry [5, Thm. 6.2.1 and Cor. 6.2.2, pp. 168–171]:

THEOREM 1. *Consider the abstract differential equation*

$$\frac{du}{dt} = Au + f(u), \tag{2.6}$$

where $-A$ is a sectorial operator in a Banach space Y, $0 \leqq \alpha < 1$, U is a neighborhood of the origin in Y^α, and f: $U \to Y$ is C^1 with $f(0) = 0$ and f' Lipschitz continuous in U. Assume $L = A + f'(0)$ has $\mathrm{Re}\,\Sigma(L) \leqq 0$ with $\Sigma(L) \cap \{\mathrm{Re}\,\lambda = 0\}$ a spectral set. Let $Y = Y_c + Y_s$ be the decomposition into L-invariant subspaces with $\mathrm{Re}\,\Sigma(L \mid_{Y_c}) = 0$ and $\mathrm{Re}\,\Sigma(L \mid_{Y_s}) < 0$.

Then there exists a C^1 local invariant manifold (a center manifold)

$$W_c = \{ u = u_c + u_s : u_s = h(u_c), u_c \in Y_c, \|u_c\|_Y < \delta \}$$

tangent to Y_c at the origin. The flow in W_c is represented by the ordinary differential equation

$$\frac{du_c}{dt} = l_c u_c + P_c g(u_c + h(u_c)), \tag{2.7}$$

where $g(u) = f(u) - f'(0)u$, $L_c = L|_{Y_c}$ and P_c is the projection of Y onto Y_c along Y_s. If the origin is asymptotically stable for (2.7), *then the origin is asymptotically stable in Y^α for* (2.6); *if the origin is unstable for* (2.7), *then the origin is unstable in Y^α for* (2.6).

If the nonlinear part f in Theorem 1 is smooth enough, the center manifold W_c is smooth and we can approximate $h(u_c)$ by the first finitely many terms in the Taylor series for $h(u_c)$ [5, Thm. 6.2.3, p. 171]:

THEOREM 2. *Assume the hypotheses of Theorem* 1, *and assume that g: $U \to Y$ is C^p, where $g(u) = f(u) + f'(0)u$. If there is a C^1 function ϕ with Lipschitzian derivative from a neighborhood of the origin in Y_c into Y_s^α, with range in $D(L|_{Y_s})$, such that*

$$\phi'(u_c)\left[L_c u_c + P_c g(u_c + \phi(u_c))\right] - L_s \phi(u_c) - P_s g(u_c + \phi(u_c)) = O\left(\|u_c\|^p\right) \tag{2.8}$$

as $u_c \to 0$ in Y_c, where $L_c = L|_{Y_c}$, $L_s = |_{Y_s}$, P_c is the projection of Y onto Y_c along Y_s, $P_s = I - P_c$, then

$$|h(u_c) - \phi(u_c)|_\alpha = O\left(\|u_c\|^p\right) \tag{2.9}$$

as $u_c \to 0$ in Y_c, where $h(u_c)$ defines the center manifold of Theorem 1. *If g is C^p near the origin, there is a unique polynomial function ϕ of order p satisfying the conditions above.*

We apply Theorems 1 and 2 to the system

$$\frac{du}{dt} = Au + f(u,r,s), \quad \frac{dr}{dt} = 0, \quad \frac{ds}{dt} = 0, \tag{2.10}$$

in the Hilbert space $Y = X \times \mathbb{R}^2$, where X is the space defined in §1, $-A$ is the sectorial operator defined by (1.11) and f is the operator defined by (1.16).

By the remarks in §1,

$$L_0 = L(r_0, s_0) = A + D_u f(0, r_0, s_0) \tag{2.11}$$

has $\mathrm{Re}\,\Sigma(L_0) \leqq 0$ with $\Sigma(L_0) \cap \{\mathrm{Re}\,\lambda = 0\} = \{0\}$. Suppose we find a Jordan basis for the zero eigenspace satisfying

$$\begin{aligned} &L_0 q_1 = 0, && L^*_0 q^*_1 = q^*_2, \\ &L_0 q_2 = q_1, && L^*_0 q^*_2 = 0, \\ &(q_j, q^*_k) = \delta_{jk}, && j,k = 1,2, \end{aligned} \tag{2.12}$$

where $(\ ,\ )$ is the inner product in X and L^*_0 is the adjoint of L_0 (see Part I). Then the operator P_c: $X \to \mathrm{span}\{q_1, q_2\}$ defined by

$$P_c u = (u, q^*_1) + (u, q^*_2) q_2 \tag{2.13}$$

is a projection onto the double zero eigenspace $\mathrm{span}\{q_1, q_2\}$, with

$$P_c L_0 u = L_0 P_c u \quad \text{for all } u \in D(L_0). \tag{2.14}$$

Thus L_0 leaves the subspaces

$$R(P_c)=\operatorname{span}\{q_1,q_2\}, \qquad N(P_c)=\{q^*_1,q^*_2\}^{\perp}$$

invariant, with

$$X=R(P_c)\oplus N(P_c) \tag{2.15}$$

and

$$\Sigma\left(L_0|_{R(P_c)}\right)=\{0\}, \qquad \operatorname{Re}\Sigma\left(L_0|_{N(P_c)}\right)<0. \tag{2.16}$$

Define

$$X_c=R(P_c), \qquad X_s=N(P_c). \tag{2.17}$$

Then by (2.15) and (2.16), X_c is the generalized eigenspace for the double zero eigenvalue of L_0 and X_s is a complementary subspace. By Theorem 1 applied to (2.10), we obtain the existence of a local center manifold

$$W_c=\left\{\left(u_c,r',s'\right)+h\left(u_c,r',s'\right)\colon \|u_c\|_X<\delta,\ |r'|<\delta,\ |s'|<\delta\right\},$$

where $r'=r-r_0$, $s'=s-s_0$ and h is a C^1 mapping from a neighborhood of the origin in $X_c\times\mathbb{R}^2$ into $X_s^\alpha=X_sX^\alpha$, $3/4<\alpha<1$, with $h(0,0,0)=0$, $h'(0,0,0)=0$. Actually, h is C^p for any integer $p>0$ (although h is not necessarily analytic) since f in (2.10) is analytic. The flow in W_c is represented by

$$\frac{du_c}{dt}=P_cL(r,s)u_c+P_cg\left(u_c+h\left(u_c,r',s'\right)\right), \tag{2.18}$$

where $g(u)=f(u,r,s)-D_uf(0,r,s)u=M(u,u)$. Bifurcation and asymptotic behavior of solutions near $u=0$, with r near r_0 and s near s_0 in the abstract differential equation (1.17) can now be studied by means of the oridinary differential equation (2.18).

We get explicit expressions for q_1, q_2, q^*_1 and q^*_2 in (2.12) by substituting the appropriate Fourier expansions into (2.12) and solving for the Fourier coefficients. For rolls, we obtain

$$q_1=\begin{bmatrix} -\left(2^{1/2}/N\right)\sin\alpha_1x_1\cos\pi x_3 \\ (1/N)\cos\alpha_1x_1\sin\pi x_3 \\ \left(2/3\pi^{2N}\right)\cos\alpha_1x_1\sin\pi x_3 \\ \left(2/3\pi^2N\tau\right)\cos\alpha_1x_1\sin\pi x_3 \end{bmatrix}, \quad q_2=\begin{bmatrix} -2^{1/2}C\sin\alpha_1x_1\cos\pi x_3 \\ C\cos\alpha_1x_1\sin\pi x_3 \\ Q_{2,4}\cos\alpha_1x_1\sin\pi x_3 \\ Q_{2,5}\cos\alpha_1x_1\sin\pi x_3 \end{bmatrix}, \tag{2.19a}$$

$$q^*_1=\begin{bmatrix} 0 \\ 0 \\ -\left(4\sigma r_0/9\pi^4\right)\cos\alpha_1x_1\sin\pi x_3 \\ \left(4\sigma s_0/9\pi^4\tau^2\right)\cos\alpha_1x_1\sin\pi x_3 \end{bmatrix}, \quad q^*_2=\begin{bmatrix} -2^{1/2}\sin\alpha_1x_1\cos\pi x_3 \\ \cos\alpha_1x_1\sin\pi x_3 \\ \left(2\sigma r_0/3\pi^2\right)\cos\alpha_1x_1\sin\pi x_3 \\ -\left(2\sigma s_0/3\pi^2\tau\right)\cos\alpha_1x_1\sin\pi x_3 \end{bmatrix},$$

where $N=2^{1/2}(1+\sigma+\tau)/\pi^2\tau$, $C=2(1+\sigma+\tau+\sigma\tau+\tau^2)/3\pi^2(1+\sigma+\tau)N\tau$, $Q_{2,4}=(2C/3\pi^2)-(4/9\pi^4N)$ and $Q_{2,5}=(2C/3\pi^2)-(4/9\pi^4N\tau)$. For rectangles and squares, we have

(2.19b)
$$q_1=\begin{bmatrix}(2i\alpha_1/\pi N)\Phi_1\\(2i\alpha_2/\pi N)\Phi_2\\(1/N)\Phi_3\\(2/3\pi^2N)\Phi_3\\(2/3\pi^2N\tau)\Phi_3\end{bmatrix},\qquad q_2=\begin{bmatrix}(2i\alpha_1C/\pi)\Phi_1\\(2i\alpha_2C/\pi)\Phi_2\\C\Phi_3\\Q_{2,4}\Phi_3\\Q_{2,5}\Phi_3\end{bmatrix},$$

$$q^*{}_1=\begin{bmatrix}0\\0\\0\\-(4\sigma r_0/9\pi^4)\Phi_3\\(4\sigma s_0/9\pi^4\tau^2)\Phi_3\end{bmatrix},\qquad q^*{}_2=\begin{bmatrix}(2i\alpha_1/\pi)\Phi_1\\(2i\alpha_2/\pi)\Phi_2\\\Phi_3\\(2\sigma r_0/3\pi^2)\Phi_3\\-(2\sigma s_0/3\pi^2\tau)\Phi_3\end{bmatrix},$$

where C, $Q_{2,4}$ and $Q_{2,5}$ are the same expressions as those for rolls, $N=16(1+\sigma+\tau)/\alpha_1\alpha_2\tau$,

$$\begin{aligned}\Phi_1&=[\omega_{1,1}-\omega_{-1,1}-\omega_{-1,-1}+\omega_{1,-1}]\cos\pi x_3,\\\Phi_2&=[\omega_{1,1}+\omega_{-1,1}+\omega_{-1,-1}+\omega_{1,-1}]\cos\pi x_3,\\\Phi_3&=[\omega_{1,1}+\omega_{-1,1}+\omega_{-1,-1}+\omega_{1,-1}]\cos\pi x_3,\\\omega_{j_1,j_2}&=e^{i(j_1\alpha_1x_1+j_2\alpha_2x_2)}.\end{aligned}$$

For hexagons we obtain the same expressions as (2.19b), except with $N=24(1+\sigma+\tau)/\alpha_1\alpha_2\tau$,

$$\begin{aligned}\Phi_1&=[\omega_{1,1}-2\omega_{-2,0}+\omega_{1,-1}-\omega_{-1,-1}+2\omega_{2,0}-\omega_{-1,1}]\cos\pi x_3,\\\Phi_2&=[\omega_{1,1}-\omega_{-1,-1}-\omega_{-1,-1}+\omega_{-1,1}]\cos\pi x_3,\\\Phi_3&=[\omega_{1,1}+\omega_{-2,0}+\omega_{-1,-1}+\omega_{-1,-1}+\omega_{2,0}+\omega_{-1,1}]\sin\pi x_3.\end{aligned}$$

If we write

(2.20)
$$u_c=y_1q_1+y_2q_2,$$

where $y_k=(u_c,q^*{}_k)\in\mathbb{R}$, $k=1,2$, we can obtain an explicit representation for $P_cL(r,s)u_c$ in (2.18) by substituting the expressions (2.19a, b) into (2.13) and computing

(2.21)
$$P_cL(r,s)=y_2q_1+(\mu_1y_1+\mu_2y_2)q_2,$$

where u_c is given by (2.20), and

$$\text{(2.22a)}\qquad \begin{aligned}\mu_1 &= \frac{\sigma\tau}{3(1+\sigma+\tau)}\left(r'-\frac{s'}{\tau}\right),\\ \mu_2 &= \frac{2\sigma\tau}{9\pi^2(1+\sigma+\tau)}\left[\frac{1+\sigma+\tau+\sigma\tau+\tau^2}{1+\sigma+\tau}\left(r'-\frac{s'}{\tau}\right)-\tau\left(r'-\frac{s'}{\tau^2}\right)\right]\end{aligned}$$

for rolls, rectangles, squares and hexagons. One notes that μ_1, μ_2 are independent linear functions of r', s', and that (2.22a) is invertible for $\sigma>0$ and $0<\tau<1$. The inverse transformation is

$$\text{(2.22b)}\qquad \begin{aligned}r' &= -\frac{3(1+\sigma+\tau)}{\sigma(1+\sigma-\mathrm{t}-\tau^2)}\left[(\sigma+\tau)\mu_1-(3\pi^2/2\tau)(1+\sigma+\tau)\mu_2\right],\\ s' &= -\frac{3(1+\sigma+\tau)}{\sigma(1+\sigma-\sigma\tau-\tau^2)}\left[(1+\sigma)\mu_1-(3\pi^2/2)(1+\sigma+\tau)\mu_2\right].\end{aligned}$$

Thus, we can write equation (2.18) as an ordinary differential equation

$$\text{(2.23)}\qquad \frac{d}{dt}\begin{bmatrix}y_1\\ y_2\end{bmatrix}=\begin{bmatrix}0 & 1\\ \mu_1 & \mu_2\end{bmatrix}\begin{bmatrix}y_1\\ y_2\end{bmatrix}+\begin{bmatrix}g_1(y,\mu)\\ g_2(y,\mu)\end{bmatrix},$$

where $g_k(y,\mu)=(M(u_c+h(u_c,r',s'),u_c+h(u_c,r',s')),q^*_k)$ for $k=1, 2$, $y=(y_1,y_2)$, $\mu=(\mu_1,\mu_2)$, $u_c=y_1q_1+y_2q_2$ as in (2.20), and r', s' are given (2.22b).

Next, we apply Theorem 2 to obtain a polynomial approximation to a center manifold. Let

$$\text{(2.24)}\qquad \phi(y,\mu)=\frac{1}{2}y_1^2\phi_{11}+y_1y_2\phi_{12}+\frac{1}{2}y_2^2\phi_{22},\qquad \phi_{jk}\in X_x^\alpha,$$

where $u_c=y_1q_1+y_2q_2$ and $\mu=(\mu_1,\mu_2)$. By substituting the expressions (2.20) and (2.24) into (2.8), and equating coefficients of powers of y_k, we obtain the equations

$$\text{(2.25)}\qquad \begin{aligned}L_s\phi_{11}\omega_{1,-1} &= -2P_sM(q_1,q_1),\\ L_s\phi_{12} &= -P_s[M(q_1,q_2)+M(q_2,q_1)]+\phi_{11},\\ L_s\phi_{22} &= -2P_sM(q_2,q_2)+2\phi_{12}.\end{aligned}$$

When the solutions ϕ_{jk} of (2.25) are substituted into (2.24), then by Theorem 2 we have the approximation

$$\text{(2.26)}\qquad h(y,\mu)=\phi(y,\mu)+O(|y||\mu|)+O\left(|(y,\mu)|^3\right)$$

as $(y,\mu)\to O$ in $\mathbb{R}^4$, where $h(y,\mu)=h(u_c,r',s')$, using (2.20) and (2.22b).

We can solve (2.25) uniquely and obtain explicit expressions for ϕ_{11}, ϕ_{12} and ϕ_{22} by substituting the expressions (2.19) for q_k into the right-hand sides of (2.25) and solving for the Fourier coefficients of the ϕ_{jk}. It turns out that $P_sM(q_j,q_k)=M(q_j,q_k)$, $j, k=1$, 2, due to the multiplication and orthogonality properties of trigonometric polynomials.

For example, for rolls

$$M(q_1,q_1)=-\frac{1}{2}\Pi_0\begin{bmatrix}(2\alpha_1/N^2)\sin 2\alpha_1 x_1\\(\pi/N^2)\sin 2\pi x_3\\(2/3\pi N^2)\sin 2\pi x_3\\(2/3\pi N^2\tau)\sin 2\pi x_3\end{bmatrix} \tag{2.27}$$

belongs to $\{q^*_1,q^*_2\}^\perp = R(P_s)$. Similar results are obtained for the other $M(q_j,q_k)$, and for the other cellular structures. An explicit expression for ϕ_{22} is not actually needed to determine the unfoldings; solving for ϕ_{11} and ϕ_{12}, we obtain, in the case of rolls,

(2.28a)
$$\phi_{11}=\begin{bmatrix}0\\0\\-(1/6\pi^3N^2)\sin 2\pi x_3\\-(1/6\pi^3N^2\tau^2)\sin 2\pi x_3\end{bmatrix},\quad \phi_{12}=\begin{bmatrix}0\\0\\\left(\dfrac{-C}{6\pi^3N}+\dfrac{7}{72\pi^5N^2}\right)\sin 2\pi x_3\\\left(\dfrac{-C}{6\pi^3N\tau^2}+\dfrac{7}{72\pi^5N^2\tau^3}\right)\sin 2\pi x_3\end{bmatrix}.$$

For rectangles and squares we obtain

$$\phi_{11}=\begin{bmatrix}2\hat\phi_{11,1}(2,0,2)\Theta_{1,1}\\2i\hat\phi_{11,3}(2,0,2)\Theta_{1,3}\\2i\hat\phi_{11,4}(2,0,2)\Theta_{1,3}\end{bmatrix}+\begin{bmatrix}0\\2\hat\phi_{11,2}(0,2,2)\Theta_{2,2}\\2i\hat\phi_{11,3}(0,2,2)\Theta_{2,3}\\2i\hat\phi_{11,4}(0,2,2)\Theta_{2,3}\\2i\hat\phi_{11,5}(0,2,2)\Theta_{2,3}\end{bmatrix} \tag{2.28b}$$

$$+\begin{bmatrix}0\\0\\0\\-(4/3\pi^3N^2)\sin 2\pi x_3\\-(4/3\pi^3N^2\tau^2)\sin 2\pi x_3\end{bmatrix},$$

$$\phi_{12}=\begin{bmatrix}2\hat\phi_{12,1}(2,0,2)\Theta_{1,1}\\2i\hat\phi_{12,3}(2,0,2)\Theta_{1,3}\\2i\hat\phi_{12,4}(2,0,2)\Theta_{1,3}\\2i\hat\phi_{12,5}(2,0,2)\Theta_{1,3}\end{bmatrix}+\begin{bmatrix}0\\2\hat\phi_{12,2}(0,2,2)\Theta_{2,2}\\2i\hat\phi_{12,3}(0,2,2)\Theta_{2,3}\\2i\hat\phi_{12,4}(0,2,2)\Theta_{2,3}\\2i\hat\phi_{12,5}(0,2,2)\Theta_{2,3}\end{bmatrix}$$

$$+\begin{bmatrix}0\\0\\0\\\left(\dfrac{-4C}{3\pi^3N}+\dfrac{7}{9\pi^5N^2}\right)\sin 2\pi x_3\\\left(\dfrac{-4C}{3\pi^3N\tau^2}+\dfrac{7}{9\pi^5N^2\tau^3}\right)\sin 2\pi x^3\end{bmatrix},$$

where

$$\Theta_{1,1}=[\omega_{2,0}-\omega_{-2,0}]\cos 2\pi x_3,$$
$$\Theta_{1,3}[\omega_{2,0}+\omega_{-2,0}]\sin 2\pi x_3,$$
$$\Theta_{2,2}=[\omega_{0,2}-\omega_{0,-2}]\cos 2\pi x_3,$$
$$\Theta_{2,3}=[\omega_{0,2}+\omega_{0,-2}]\sin 2\pi x_3,$$

$$\hat{\phi}_{11,1}(2,0,2)=-(\pi/\alpha_1)\hat{\phi}_{11,3}(2,0,2),$$

$$\hat{\phi}_{11,3}(2,0,2)=2i\alpha_1^2\delta_1^2/\pi\gamma_1^6N^2\sigma,$$

$$\hat{\phi}_{11,4}(2,0,2)=\left[1/4\left(\alpha_1^2+\pi^2\right)\right]\left[\hat{\phi}_{11,3}(2,0,2)+\left(2i\delta_1^2/3\pi^5N^2\right)\right],$$

$$\hat{\phi}_{11,5}(2,0,2)=\left[1/4\left(\alpha_1^2+\pi^2\right)\tau\right]\left[\hat{\phi}_{11,3}(2,0,2)+\left(2i\delta_1^2/3\pi^5N^2\tau\right)\right],$$

$$\hat{\phi}_{11,2}(0,2,2)=-(\pi/\alpha_2)\hat{\phi}_{11,3}(0,2,2),$$

$$\hat{\phi}_{11,3}(0,2,2)=2i\alpha_2^2\delta_2^2/\pi\gamma_2^6N^2\sigma,$$

$$\hat{\phi}_{11,4}(0,2,2)=\left[1/4\left(\alpha_2^2+\pi^2\right)\right]\left[\hat{\phi}_{11,3}(0,2,2)+\left(2i\delta_2^2/3\pi^5N^2\right)\right],$$

$$\hat{\phi}_{11,5}(0,2,2)=\left[1/4\left(\alpha_2^2+\pi^2\right)\tau\right]\left[\hat{\phi}_{11,3}(0,2,2)+\left(2i\delta_2^2/3\pi^5N^2\tau\right)\right],$$

$$\hat{\phi}_{12,1}(2,0,2)=-(\pi/\alpha_1)\hat{\phi}_{12,3}(2,0,2),$$

$$\hat{\phi}_{12,3}(2,0,2)=-(1/4\sigma)\hat{\phi}_{11,3}(2,0,2)$$
$$+\left(8i\alpha_1^2\delta_1^2/\gamma_1^6\right)\left[\frac{C\left(24\alpha_1^2+15\pi^2\right)}{4\pi N}-\frac{\left(4\alpha_1^2+7\pi^2\right)\sigma}{36\pi^5\left(\alpha_1^2+\pi^2\right)N^2}\left(r_0-\frac{s_0}{\tau^3}\right)\right],$$

$$\hat{\phi}_{12,4}(2,0,2)=\hat{\phi}_{12,3}(2,0,2)/4\left(\alpha_1^2+\pi^2\right)-\hat{\phi}_{11,3}(2,0,2)/16\left(\alpha_1^2+\pi^2\right)^2$$
$$+\frac{i\delta_1^2}{6\pi^3\left(\alpha_1+\pi^2\right)}\left[\frac{C}{N}-\frac{4\alpha_1^2+7\pi^2}{12\left(\alpha_1^2+\pi^2\right)N^2}\right],$$

$$\hat{\phi}_{12,5}(2,0,2)=\hat{\phi}_{12,3}(2,0,2)/4\left(\alpha_1^2+\pi^2\right)\tau-\hat{\phi}_{11,3}(2,0,2)/16\left(\alpha_1^2+\pi^2\right)^2\tau^2$$
$$+\frac{i\delta_1^2}{6\pi^3\left(\alpha_1+\pi^2\right)\tau}\left[\frac{C}{N}-\frac{4\alpha_1^2+7\pi^2}{12\left(\alpha_1^2+\pi^2\right)N^2\tau}\right],$$

$$\hat{\phi}_{12,2}(0,2,2)=-(\pi/\alpha_2)\hat{\phi}_{12,3}(0,2,2),$$

$$\hat{\phi}_{12,3}(0,2,2)=-(1/4\sigma)\hat{\phi}_{11,3}(0,2,2)$$
$$+\left(8i\alpha_2^2\delta_2^2/\gamma_2^6\right)\left[\frac{C\left(24\alpha_2^2+15\pi^2\right)}{4\pi N}-\frac{\left(4\alpha_2^2+7\pi^2\right)\sigma}{36\pi^5\left(\alpha_2^2+\pi^2\right)N^2}\left(r_0-\frac{s_0}{\tau^3}\right)\right],$$

$$\hat{\phi}_{12,4}(0,2,2)=\hat{\phi}_{12,3}(0,2,2)/4\left(\alpha_2^2+\pi^2\right)-\hat{\phi}_{11,3}(0,2,2)/16\left(\alpha_2^2+\pi^2\right)^2$$
$$+\frac{i\delta_2^2}{6\pi^3\left(\alpha_2^2+\pi^2\right)}\left[\frac{C}{N}-\frac{4\alpha_2^2+7\pi^2}{12\left(\alpha_2^2+\pi^2\right)N^2}\right],$$

$$\hat{\phi}_{12,5}(0,2,2)=\hat{\phi}_{12,3}(0,2,2)/4\left(\alpha_2^2+\pi^2\right)\tau-\hat{\phi}_{11,3}(0,2,2)/16\left(\alpha_2^2+\pi^2\right)^2\tau^2$$
$$+\frac{i\delta_2^2}{6\pi^3\left(\alpha_2^2+\pi^2\right)\tau}\left[\frac{C}{N}-\frac{4\alpha_2^2+7\pi^2}{12\left(\alpha_2^2+\pi^2\right)N^2\tau}\right],$$
$$\delta_k^2=\pi+(-1)^k 2\left(\alpha_1^2-\alpha_2^2\right),\qquad k=1,2,$$
$$\gamma_k^6=64\left(\alpha_k^2+\pi^2\right)^3-27\alpha_k^2\pi^4,\qquad k=1,2.$$

For hexagons we obtain

(2.28c)
$$\phi_{11}=\begin{bmatrix}2\hat{\phi}_{11,1}(1,1,2)\Phi_1\\2\hat{\phi}_{11,2}(1,1,2)\Phi_2\\2i\hat{\phi}_{11,3}(1,1,2)\Phi_3\\2i\hat{\phi}_{11,4}(1,1,2)\Phi_3\\2i\hat{\phi}_{11,5}(1,1,2)\Phi_3\end{bmatrix}\begin{bmatrix}2\hat{\phi}_{11,1}(3,1,2)\Theta_1\\2\hat{\phi}_{11,2}(3,1,2)\Theta_2\\2i\hat{\phi}_{11,3}(3,1,2)\Theta_3\\2i\hat{\phi}_{11,4}(3,1,2)\Theta_3\\2i\hat{\phi}_{11,4}(3,1,2)\Theta_3\end{bmatrix}$$
$$+\begin{bmatrix}0\\0\\0\\-(2/\pi^3N^2)\sin 2\pi x_3\\-(2/\pi^3N^2\tau^2)\sin 2\pi x_3\end{bmatrix},$$

$$\phi_{12}=\begin{bmatrix}2\hat{\phi}_{12,1}(1,1,2)\Phi_1\\2\hat{\phi}_{12,2}(1,1,2)\Phi_2\\2i\hat{\phi}_{12,3}(1,1,2)\Phi_3\\2i\hat{\phi}_{12,4}(1,1,2)\Phi_3\\2i\hat{\phi}_{12,5}(1,1,2)\Phi_3\end{bmatrix}+\begin{bmatrix}2\hat{\phi}_{12,1}(3,1,2)\Theta_1\\2\hat{\phi}_{12,2}(3,1,2)\Theta_2\\2i\hat{\phi}_{12,3}(3,1,2)\Theta_3\\2i\hat{\phi}_{12,4}(3,1,2)\Theta_3\\2i\hat{\phi}_{12,5}(3,1,2)\Theta_3\end{bmatrix}$$
$$+\begin{bmatrix}0\\0\\0\\\left(\frac{-2C}{\pi^3N}+\frac{1}{2\pi^5N^2}\right)\sin 2\pi x_3\\\left(\frac{-2C}{\pi^3N\tau^2}+\frac{1}{2\pi^5N^2\tau^3}\right)\sin 2\pi x_3\end{bmatrix},$$

where

$$\Phi_1 = [\omega_{1,1} - 2\omega_{-2,0} + \omega_{1,-1} - \omega_{-1,-1} + 2\omega_{2,0} - \omega_{-1,1}]\cos 2\pi x_3,$$
$$\Phi_2 = [\omega_{1,1} - \omega_{1,-1} - \omega_{-1,-1} + \omega_{-1,1}]\cos 2\pi x_3,$$
$$\Phi_2 = [\omega_{1,1} + \omega_{-2,0} + \omega_{1,-1} + \omega_{-1,-1} + \omega_{2,0} + \omega_{-1,1}]\sin 2\pi x_3,$$
$$\Theta_1 = [3\omega_{3,1} - 3\omega_{-3,1} - 3\omega_{-3,-1} + \omega_{3,-1}]\cos 2\pi x_3,$$
$$\Theta_2 = [\omega_{3,1} + \omega_{-3,1} - 2\omega_{0,-2} - \omega_{-3,-1} - \omega_{3,-1} + 2\omega_{0,2}]\cos 2\pi x_3,$$
$$\Theta_3 = [\omega_{3,1} + \omega_{-3,1} - \omega_{0,-2} + \omega_{-3,-1} + \omega_{3,-1} + \omega_{0,2}]\sin 2\pi x_3,$$
$$\hat{\phi}_{11,1}(1,1,2) = -2^{1/2}\hat{\phi}_{11,3}(1,1,2),$$
$$\hat{\phi}_{11,2}(1,1,2) = -6^{1/2}\hat{\phi}_{11,3}(1,1,2),$$
$$\hat{\phi}_{11,3}(1,1,2) = i/13\pi N^2\sigma,$$
$$\hat{\phi}_{11,4}(1,1,2) = (2i/9\pi^3 N^2)[1 + (1/13\sigma)],$$
$$\hat{\phi}_{11,5}(1,1,2) = (2i/9\pi^3 N^2)[(1/\tau^2) + (1/13\sigma\tau)],$$
$$\hat{\phi}_{11,1}(3,1,2) = -(2^{1/2}/3)\hat{\phi}_{11,3}(3,1,2),$$
$$\hat{\phi}_{11,2}(3,1,2) = -(6^{1/2}/3)\hat{\phi}_{11,3}(3,1,2),$$
$$\hat{\phi}_{11,3}(3,1,2) = 36i/625\pi N^2\sigma,$$
$$\hat{\phi}_{11,4}(3,1,2) = (2i/\pi^3 N^2)[(1/33) + (36/6,875\sigma)],$$
$$\hat{\phi}_{11,5}(3,1,2) = (2i/\pi^3 N^2)[(1/33\tau^2) + (36/6,875\sigma\tau)],$$
$$\hat{\phi}_{12,1}(1,1,2) = -2^{1/2}\hat{\phi}_{12,3}(1,1,2),$$
$$\hat{\phi}_{12,2}(1,1,2) = -6^{1/2}\hat{\phi}_{12,3}(1,1,2),$$
$$\hat{\phi}_{12,3}(1,1,2) = \hat{\phi}_{11,3}(1,1,2)/117\pi^2\sigma + iC/13\pi N\sigma - i/338\pi^3 N^2\sigma^2 - (11i/6,318\pi^7 N^2)[r_0 - (s_0/\tau^3)],$$
$$\hat{\phi}_{12,4}(1,1,2) = 2\hat{\phi}_{12,3}(1,1,2)/9\pi^2 - 4\hat{\phi}_{11,3}(1,1,2)/81\pi^4 + 2iC/9\pi^3 N - 11i/162\pi^5 N^2,$$
$$\hat{\phi}_{12,5}(1,1,2) = 2\hat{\phi}_{12,3}(1,1,2)/9\pi^2\tau - 4\hat{\phi}_{11,3}(1,1,2)/81\pi^4\tau^2 + 21C/9\pi^3 N\tau^2 - 11i/162\pi^5 N^2\tau^3,$$
$$\hat{\phi}_{12,1}(3,1,2) = -(2^{1/2}/3)\hat{\phi}_{12,3}(3,1,2),$$
$$\hat{\phi}_{12,2}(3,1,2) = -(6^{1/2}/3)\hat{\phi}_{12,3}(3,1,2),$$
$$\hat{\phi}_{12,3}(3,1,2) = 81\hat{\phi}_{11,3}(3,1,2)/6,875\pi^2\sigma + 36iC/675\pi N\sigma - 4,356i/390,625\pi^3 N^2\sigma^2 - 7i/1,650\pi^7 N^2[r_0 - (s_0/\tau^3)],$$

$$\hat{\phi}_{12,4}(3,1,2)=2\hat{\phi}_{12,3}(3,1,2)/11\pi^2-4\hat{\phi}_{11,3}(3,1,2)/121\pi^4$$
$$+2iC/33\pi^3N-35i/2{,}178\pi^5N^2,$$
$$\hat{\phi}_{12,5}(3,1,2)=2\hat{\phi}_{12,3}(3,1,2)/11\pi^2\tau-4\hat{\phi}_{11,3}(3,1,2)/121\pi^4\tau^2$$
$$+2iC/33\pi^3N\tau^2-35i/2{,}178\pi^5N^2\tau^3.$$

If we substitute (2.20), (2.24) and (2.26) into equation (2.23), we obtain

$$\frac{dy_1}{dt}=y_2+a_1y_1^3+b_1y_1^2y_2+c_1y_1y_2^2+d_1y_2^3+\rho_1(y,\mu), \tag{2.29}$$
$$\frac{dy_2}{dt}=\mu_1y_1+\mu_2y_2+a_2y_1^3+b^2y_1^2y_2+c_2y_1y_2^2+d_2y_2^3+\rho_2(y,\mu).$$

where $\rho_k(y,\mu)=0(|y||\mu|^2+|y|^2|\mu|)+0(|(y,\mu)|^4)$, and

$$a_k=\frac{1}{2}(M(q_1,\phi_{11})+M(\phi_{11},q_1),q^*{}_k), \tag{2.30}$$
$$b_k=(M(q_1,\phi_{12})+M(\phi_{12},q_1)q^*{}_k)+\frac{1}{2}(M(q_2,\phi_{11})+M(\phi_{11},q_2),q^*{}_k),$$
$$c_k=(M(q_2,\phi_{12})+M(\phi_{12},q_2),q^*{}_k)+\frac{1}{2}(M(q_1,\phi_{22})+M(\phi_{22},q_1),q^*{}_k),$$
$$d_k=\frac{1}{2}(M(q_2,\phi_{22})+M(\phi_{22},q_2),q^*{}_k),$$

for $k=1, 2$. We have used the fact that $M(q_j,q_k)\in\{q_1,q_2\}^{\perp}$, $j, k=1, 2$ so that the quadratic terms in y vanish. In fact, since the functions in X^{α} and equation (1.17) are covariant with respect to the $\mathbb{Z}_2$ symmetry represented by (1.4), the reduced equations (2.29) are odd with respect to y. Thus, no even-order terms in y occur in (2.29) [4, p. 256] and we have in fact

$$\rho_k(y,\mu)=0\big(|y||\mu|^2+|y|^3|\mu|+|y|^5\big),\qquad k=1,2.$$

In the following section we determine the unfolding of the degenerate vector field represented by the right-hand side of (2.29) about $\mu=0$, and hence determine the unfolding of the parametrized family of abstract differential equations (1.17) about $(r,s)=(r_0,s_0)$ for almost all $\sigma>0$ and $0<\tau<1$.

3. Normal form coefficients and results. The flow of a nonlinear ordinary differential equation near a degenerate equilibrium point can be analyzed, after a center manifold reduction, by means of a nonlinear change of coordinates which simplifies the expression of the vector field. A normal form is a "simplest" expression resulting from a smooth nonlinear change of coordinates. More precisely, for certain differential equations the first finitely many terms in the Taylor series expansion of the vector field are sufficient to determine the asymptotic behavior of solutions near a degenerate equilibrium. In such cases, the normal form theorem ([11], see also [6, p. 459]) is used to find the minimum number of terms essential to describe the local flow. One can then construct coordinate transformations which transform the original vector field into a normal form, modulo higher-order terms. This procedure can be modified for use with parametrized systems such as (2.29) to determine the unfolding of a degenerate vector field [6, pp. 473–474].

In this section we present the results of our computations of the normal form coefficients which determine the unfolding of (2.29). The values of these coefficients completely determine all the local bifurcations which occur in (2.29) for μ near 0, as well as the stability of each of the bifurcating solutions. In this way we obtain complete information about local bifurcation and stability in the abstract equations (1.17) for (r,s) near (r_0,s_0), for almost all values of $\sigma>0$ and $0<\tau<1$.

For equation (2.29), terms of order three in y are sufficient to determine the flow for (y,μ) near $(0,0)$, and a suitable normal form is

$$\frac{dy_1}{dt}=y_2, \qquad \frac{dy_2}{dt}=\mu_1 y_1+\mu_2 y_2+a_2' y_1^3+b_2' y_1^2 y_2, \tag{3.1}$$

if $a_2'\neq 0$ and $b_2'\neq 0$ [6, p. 474]. Since two parameters μ_1 and μ_2 are necessary and sufficient to determine all the local bifurcations in (3.1), this is called a codimension two bifurcation. By a suitable smooth change of coordinates preserving the symmetry $y\to -y$, we can transform (2.29) into

$$\frac{dy_1}{dt}=y_2+\rho_1(y,\mu), \qquad \frac{dy_2}{dt}=\mu_1 y_1+\mu_2 y_2+a_2' y_1^3+b_2' y_1^2 y_2+\rho_2(y,\mu), \tag{3.2}$$

where $\rho_k(y,\mu)=O(|y||\mu|^2+|y|^3|\mu|+|y|^3)$ for $k=1, 2$, and the normal form coefficients are

$$a_2'=a_2, \qquad b_2'=3a_1+b_2 \tag{3.3}$$

[6, p. 466], [1, p. 81]. The values of a_2' and b_2' completely determine the qualitative behaviors of all the local flows of (3.2), and hence of (2.29) for μ near 0.

We compute the values of the normal form coefficients by substituting expressions (2.19), (2.28) and (2.30) into (3.3). For rolls, we obtain

$$a_2'=\frac{9}{2^{1/2}8N^3}\left[\frac{1+\sigma+\tau}{\tau}\right], \qquad b_2'=-\frac{21}{2^{1/2}16\pi^2N^3}\left[\frac{1+\sigma+\tau+\sigma\tau+\tau^2}{\tau^2}\right]. \tag{3.4a}$$

For rectangles, we obtain

$$\begin{aligned} a_2'=\frac{8\pi^2}{\alpha_1\alpha_2 N_3}\Bigg\{&\frac{1+\sigma+\tau}{\tau}\left[4+\sum_{k=1}^{2}\frac{(\pi^2-2\alpha_k^2)\delta_k^2}{\pi^2(\pi^2+\alpha_k^2)}\right]\\ &-\frac{1}{\sigma}\left[\sum_{k=1}^{2}\frac{3\alpha_k^2(\pi^2-2\alpha_k^2)(5\pi^2+8\alpha_k^2)^2\delta_k^2}{\pi^2(\pi^2+\alpha_k^2)\gamma^6}\right]\Bigg\}, \end{aligned}$$

$$\begin{aligned} b_2'=\frac{8}{\alpha_1\alpha_2 N^3}\Bigg\{&-\frac{1+\sigma+\tau+\sigma\tau+\tau^2}{\tau^2}\left[7+\sum_{k=1}^{2}\frac{(7\pi^4-10\pi^2\alpha_k^2-8\alpha_k^4)\delta_k^2}{4\pi^2(\pi^2+\alpha_k^2)}\right]\\ &-\frac{1+\sigma+\tau+\sigma\tau+\tau^2}{\sigma\tau(1+\sigma+\tau)}\left[\sum_{k=1}^{2}\frac{7\alpha_k^2(\pi^2-2\alpha_k^2)(5\pi^2+8\alpha_k^2)^2\delta_k^2}{\pi^2(\pi^2+\alpha_k^2)\gamma_k^6}\right]\\ &-\frac{1+\sigma+\tau}{\sigma\tau}\left[\sum_{k=1}^{2}\frac{36\alpha_k^2(\pi^2-2\alpha_k^2)(5\pi^2+8\alpha_k^2)\delta_k^2}{(\pi^2+\alpha_k^2)\gamma_k^6}\right]\\ &+\frac{1}{\sigma^2}\left[\sum_{k=1}^{2}\frac{9\alpha_k^2(\pi^2-2\alpha_k^2)(5\pi^2+8\alpha_k^2)^2\delta_k^2}{(\pi^2+\alpha_k^2)^2\gamma_k^6}\right]\Bigg\}. \end{aligned} \tag{3.4b}$$

Note that the quantitites in square brackets are positive. For squares, the expressions (3.4b) simplify to

$$a_2' = \frac{32}{N^3}\left\{\frac{36}{5}\left[\frac{1+\sigma+\tau}{\tau}\right] - \frac{882}{2{,}365}\left[\frac{1}{\sigma}\right]\right\},$$

$$\text{(3.4c)}\quad b_2' = \frac{32}{\pi^2 N^3}\left\{-\frac{207}{25}\left[\frac{1+\sigma+\tau+\sigma\tau+\tau^2}{\tau^2}\right] - \frac{2{,}352}{2{,}365}\left[\frac{1+\sigma+\tau+\sigma\tau+\tau^2}{\sigma\tau(1+\sigma+\tau)}\right] - \frac{18{,}816}{11{,}875}\left[\frac{1+\sigma+\tau}{\sigma\tau}\right] + \frac{441}{2{,}365}\left[\frac{1}{\sigma^2}\right]\right\}.$$

For hexagons, we obtain

$$a_2' = \frac{48}{3^{1/2}N^3}\left\{\frac{810}{11}\left[\frac{1+\sigma+\tau}{\tau}\right] - \frac{314{,}892}{89{,}375}\left[\frac{1}{\sigma}\right]\right\},$$

$$\text{(3.4d)}\quad b_2' = \frac{48}{3^{1/2}\pi^2 N^3}\left\{-\frac{226}{33}\left[\frac{1+\sigma+\tau+\sigma\tau+\tau^2}{\tau^2}\right] - \frac{242{,}928}{89{,}375}\left[\frac{1+\sigma+\tau+\sigma\tau+\tau^2}{\sigma\tau(1+\sigma+\tau)}\right] - \frac{1{,}092{,}828}{983{,}125}\left[\frac{1+\sigma+\tau}{\sigma\tau}\right] + \frac{471{,}269{,}040}{726{,}171{,}875}\left[\frac{1}{\sigma^2}\right]\right\}.$$

The bifurcation structure of (3.2) is now a straightforward application of the theory already developed for (3.1). The unfolding of (3.2) is completely determined by the signs of the coefficients a_2' and b_2'. The following cases are possible [1, Chap. 4], [3, Chap. 7]:

Case 1a. $a_2' > 0$, $b_2' < 0$: See Fig. 3.1 for bifurcation set and associated phase portraits.

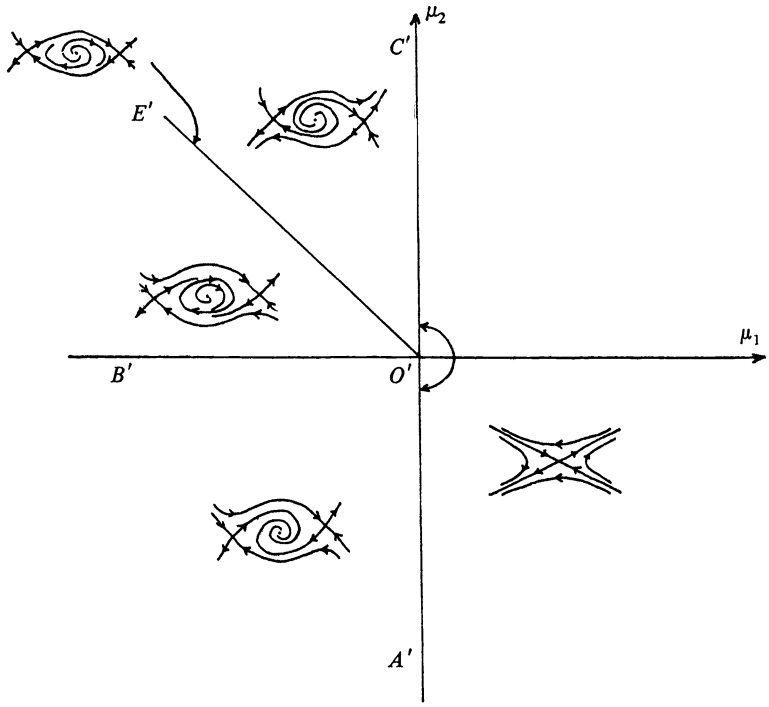


FIG. 3.1. *Bifurcation set and corresponding phase portraits for Case* 1a *unfolding* ($a_2' > 0$, $b_2' < 0$).

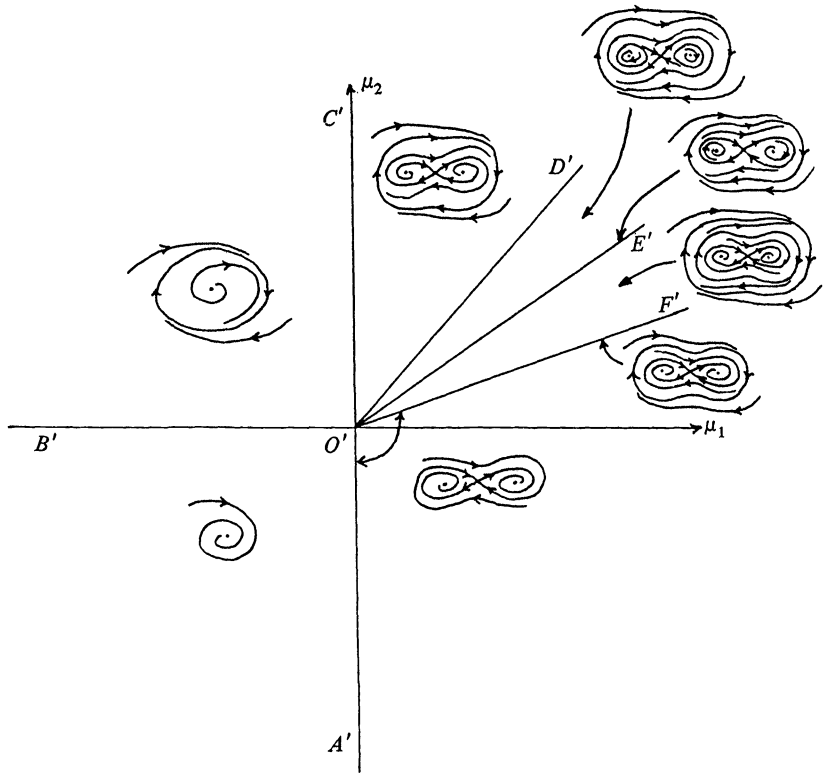


FIG. 3.2. *Bifurcation set and corresponding phase portraits for Case* 2a *unfolding* ($a_2' < 0$, $b_2' < 0$).

Case 1b. $a_2' > 0$, $b_2' > 0$: Bifurcation set and associated phase portraits for this case are obtained from Fig. 3.1 by the transformations $\mu_2 \to -\mu_2$, $y_2 \to -y_2$ and $t \to -t$.

Case 2a. $a_2' < 0$, $b_2' < 0$: See Fig. 3.2 for bifurcation and set associated phase portraits.

Case 2b. $a_2' < 0$, $b_2' > 0$: Bifurcation and associated phase portraits for this case are obtained from Fig. 3.2 by the transformations $\mu_2 \to -\mu_2$, $y_2 \to -y_2$ and $t \to -t$.

The lines $A'O'C'$ and $O'B'$ in (μ_1, μ_2)-parameter space shown in Figs. 3.1–3.2, correspond to the lines AOC and OB in (s, r)-parameter space shown in Figs. 3.3 and 3.5. The corresponding bifurcations—pitchfork and Hopf, respectively—are local ones branching from the trivial (constant gradient) solution $u = 0$ of (1.17) when $s \neq s_0$. They are treated in Parts I and II.

However, we have further local bifurcations. The line $O'D'$ in Fig. 3.2 corresponds to Hopf bifurcations from two equilibrium points symmetrically located about the origin in phase space. The curves $O'E'$ and $O'F'$ in Figs. 3.1 and 3.2 are tangent to straight lines at the origin in parameter space, and correspond to saddle connections and coalescence of xperiodic orbits. For more details consult [1] or [3].

It is clear from (3.4a) that for rolls, only Case 1a unfoldings occur for all $\sigma > 0$ and $0 < \tau < 1$. However, different cases of unfoldings can occur for rectangles, squares and hexagons. For example, from (3.4c, d) it follows that Case 1b unfoldings occur for squares and hexagons when $\sigma \ll 1$ and and $\tau = 17\sigma \ll 1$, and Case 2a occurs when $\sigma \ll 1$ and $\tau = 25\sigma \ll 1$. Cases 1b and 2a can also occur for rectangles. From (3.4b, c, d) one can see that Case 2b unfoldings occur for rectangles, squares and hexagons provided σ

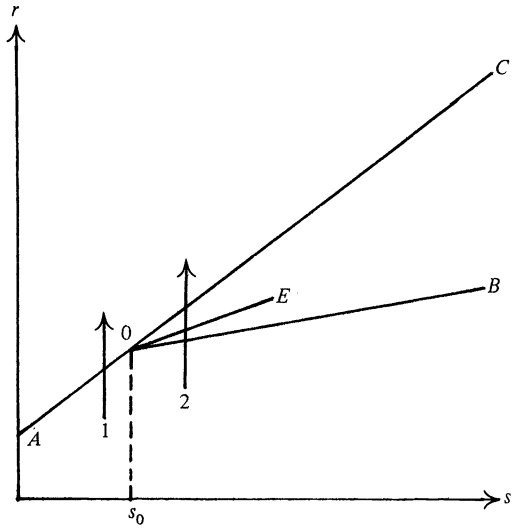


FIG. 3.3. *Case* 1a *bifurcation set in* (s,r) *parameter space.*

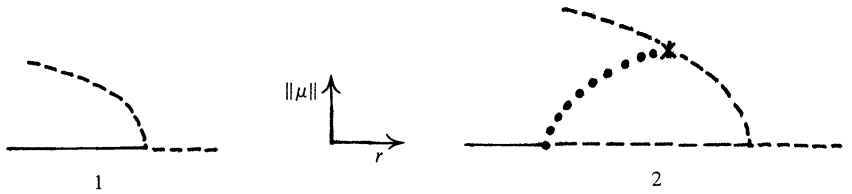


FIG. 3.4. *Bifurcation diagrams corresponding to parameter paths* 1 *and* 2 *in Fig.* 3.3. *Solid lines represent stable equilibrium solutions, broken lines represent unstable equilibrium solutions, solid circles represent stable periodic solutions, open circles represent unstable periodic solutions and asterisks represent saddle connections.*

is sufficiently small compared to τ, and the Case 1a unfoldings occur when σ is sufficiently large compared to τ.

The bifurcation sets in terms of the original parameters r and s of the double-diffusive convection equations can be found by means of (2.22a, b). The slopes of the tangent lines to the curves $O'D'$, $O'E'$ and $O'F'$ in Figs. 3.1–3.2 depend on the numerical values of a_2' and b_2', but their relative positions with respect to each other and to the lines $A'O'C'$ and $O'B'$ are unchanged. Thus, the corresponding tangent lines to the curves OD, OE and OF in (s,r) parameter space maintain their relative positions with respect to each other and to the lines AOC and OB. For example, in Case 1a we have the bifurcation set illustrated in Fig. 3.3, valid for r near r_0 and s near s_0. Following parameter path 1 in Fig. 3.3 (s fixed near s_0, $s<s_0$, r increasing) gives a subcritical pitchfork bifurcation, and following parameter path 2 in Fig. 3.3 gives the more complicated bifurcation diagram illustrated in Fig. 3.4. In case 2b, parameter paths 3 and 4 in Fig. 3.5 correspond to the bifurcation diagrams of Fig. 3.6. It is possible that the unstable subcritical branches of equilibrium solutions in Figs. 3.4 and 3.6 may "turn back" and regain stability as they have been shown to do in the case of rolls for s near $s_1=27\pi^4\tau^3/4(1-\tau^2)$ [10]. However, we do not address this question here.

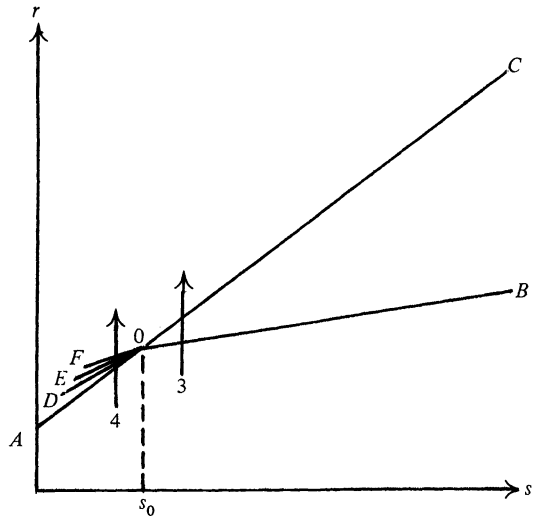


FIG. 3.5. *Case* 2b *bifurcation set in* (s,r) *parameter space.*

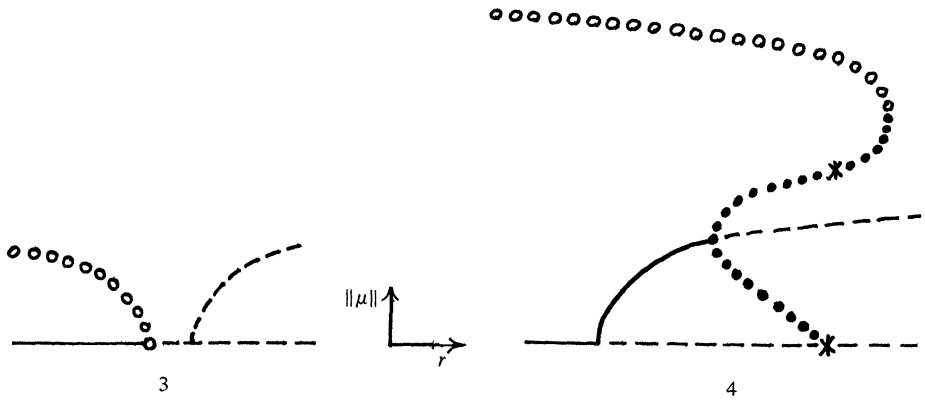


FIG. 3.6. *Bifurcation diagrams corresponding to parameter paths* 3 *and* 4 *in Fig.* 3.5. (*See Fig.* 3.4 *legend*).

"Physically reasonable" parameter values are σ near 1 or greater, and $\tau \ll 1$. These correspond to Case 1a unfoldings for all four types of cellular convection patterns considered. However, the possibility exists, at least mathematically, that rectangles, squares and hexagons can have a different unfolding than the one that rolls must have. This is implied by our previous results in Part I and II.

For parameter values σ and τ such that $a_2'=0$ or $b_2'=0$, the normal form (3.1) is inadequate to determine the unfolding of (1.17) about (r_0,s_0), but such values constitute a set of measure zero in the space of parameter values $\sigma>0$, $0<\tau<1$.

4. Conclusion. We have studied bifurcations which occur in a system of equations describing double-diffusive convection in a layer of fluid. The partial differential equations were expressed as a single abstract evolution equation in an infinite-dimensional Hilbert space corresponding to two-dimensional roll-like convection cells, or three-dimensional convection cells with rectangular, square, or hexagonal plan-forms. At the critical parameter values r_0 and s_0 for which the linear part of the equation has a

double zero eigenvalue, the infinite-dimensional equation was reduced to a finite dimensional one by means of center manifold theorem. The resulting equation was then further reduced to a normal form, from which one obtains a complete description of the local bifurcations which occur in the double-diffusive convection equations, for almost all $\sigma < 0$ and $0 < \tau < 1$. The bifurcations take place in a phase space that corresponds to a single type of cellular convection pattern. We have found that the three-dimensional convection patterns can have different bifurcations from those associated with two-dimensional roll-like convection.

Our result for rolls agrees with that of [7], who applied perturbation methods to a system of ordinary differential equations obtained from the partial differential equations by modal truncation. In addition, our result was obtained via center manifold and normal form reductions, and so no other bifurcation behavior occurs locally. One cannot conclude this from the perturbation methods alone. Numerical studies of the same ordinary differential equations show that the bifurcation results obtained locally actually extend to parameter values some distance away from the critical ones which give the double zero eigenvalue [2]. A related problem of thermally driven convection in a rotating fluid layer was treated by [4]. Center manifold and normal form reductions were applied to a similar system of ordinary differential equations resulting from modal truncation. The normal form used was the same as (3.1), and the same case of unfoldings were present. All three papers [7], [2] and [4] treated only two-dimensional roll-like convection, and in all three papers a system of partial differential equations was first reduced to a system of ordinary differential equations by means of a modal truncation.

REFERENCES

[1] J. CARR, *Applications of Centre Manifold Theory*, Springer-Verlag, New York, 1982.

[2] L. N. DA COSTA, E. KNOBLOCH AND N. O. WEISS, *Oscillations in double-diffusive convection*, J. Fluid Mech., 109 (1981), pp. 25–43.

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

[4] J. GUCKENHEIMER AND E. KNOBLOCH, *Nonlinear convection in a rotating layer: amplitude expansions and normal forms*, Geophys. Astrophys. Fluid Dynamics, 23 (1983), pp. 247–272.

[5] D. HENRY, *Geometric Theory of Semilinear Parabolic Equations*, Lecture Notes in Mathematics 840, Springer-Verlag, Berlin, 1981.

[6] P. HOLMES, *Center manifolds, normal forms and bifurcations of vector fields with applications to coupling between periodic and steady motions*, Physica, 2D (1981), pp. 449–481.

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

[8] W. NAGATA AND J. THOMAS, *Bifurcations in doubly-diffusive systems* I. *Equilibrium solutions*, this Journal, 17 (1986), pp. 91–113.

[9] ———, *Bifurcations in doubly-diffusive systems* II. *Time periodic solutions*, this Journal, 17 (1986), pp. 114–127.

[10] J. NEU, *Convective flow with subcritical instability*, Phys. Fluids, 25 (1982), pp. 8–13.

[11] F. TAKENS, *Forced oscillations and bifurcations*, Communication 3, Mathematical Institute, Rijksuniversiteit Utrecht, the Netherlands, 1974, pp. 1–59.

[12] R. TEMAM *Navier–Stokes Equations and Nonlinear Functional Analysis*, CBMS Regional Conference Series in Applied Mathematics, Society for Industrial and Applied Mathematics, Philadelphia, 1983.

[13] F. H. BUSSE, *Non-linear properties of thermal convection*, Rep. Prog. Phys., 41 (1978), pp. 1929–1967.

[14] P. G. DANIELS, *Roll-pattern evolution in finite-amplitude Rayleigh–Bénard convection in a two-dimensional fluid layer bounded by distant sidewalls*, J. Fluid Mech., 43 (1984), pp. 125–152.

[15] M. GOLUBITSKY, J. W. SWIFT AND E. KNOBLOCH, *Symmetries and pattern selection in Rayleigh–Bénard convection*, Physica, 10D (1984), pp. 249–276.

[16] D. H. SATTINGER, *Group Theoretic Methods in Bifurcation Theory*, Springer-Verlag, Berlin, 1979.

SIAM J. MATH. ANAL.
Vol. 17, No. 2, March 1986

003

SECONDARY BIFURCATIONS OF A THIN ROD UNDER AXIAL COMPRESSION*

ERNESTO BUZANO†

Abstract. We study the post-buckling behavior of a prismatic rod with rectangular cross-section, under axial compression. We employ the nonlinear rod theory stated in [2]. Let δ be the difference between the sides of the cross-section. When $\delta \neq 0$ there are two distinct eigenvalues which coalesce when $\delta = 0$ and the cross-section becomes a square. By employing equivariant singularity theory [4], we unfold the bifurcation problem corresponding to $\delta = 0$ and study also the case $\delta \neq 0$. We obtain bifurcation diagrams with second bifurcations when $\delta \neq 0$.

Introduction. A fruitful way of studying secondary bifurcation problems is that of forcing two eigenvalues to coincide by varying some parameter of the problem and then unfolding the multiple eigenvalue so obtained. This technique, due to Bauer, Keller and Reiss [11], has been coupled with singularity theory by Schaeffer and Golubitsky in [9], where they give an explanation of the phenomenon of mode-jumping observed experimentally in the post-buckling behavior of a rectangular plate under compression. Other applications are given in [3] and [10].

Here we apply this procedure to the study of secondary bifurcations of a prismatic rod with rectangular cross-section. Let δ be the difference between the sides of the cross-section. When $\delta \neq 0$ there are two distinct eigenvalues which coalesce when $\delta = 0$ and the cross-section becomes a square. In this last case we can apply the results obtained in [2], where the bifurcation problem has been reduced to finite dimension. The problem obtained in this way has topological codimension 1 in the module of the mappings which commute with the symmetries of the rectangle. Therefore it is possible to unfold it and study also the case $\delta \neq 0$. The bifurcation diagrams we obtain are given at the end of the paper. They show that secondary bifurcations occur when $\delta \neq 0$.

The classical rod model, based on linear constitutive equations (see for example [6] or [7]), has infinite codimension in our context. We reduce our problem to finite codimension by employing the nonlinear rod theory stated in [2].

Evidence for secondary bifurcations of a prismatic rod has also been obtained by Kovari in [5], by using elliptic functions and numerical computations. However these bifurcations do not coincide with ours because they concern the case $\delta = 0$ where we do not obtain secondary bifurcations.

1. Geometry of the deformation. In [2, Chap. 1] a simple nonlinear director theory has been employed to understand the effects of the symmetry of the cross-sections on the post-buckling behavior of a thin rod under axial compression. This theory is a special case of a more general situation studied in [1], but it relies upon a weaker transverse isotropy condition in order to distinguish between rods with polygonal and circular cross-section. Here we recall briefly such a theory in the case of a rectangular cross-section. The reader should refer to [2, Chap. 1] for the details.

We denote by $\mathscr{C}$ the *axis* of the rod, that is the line of centroids of the cross-sections. We assume that $\mathscr{C}$ is of length 1 and parametrize $\mathscr{C}$ by arc-length $s \in [0,1]$. Throughout the paper we set $J = [0,1]$ and denote by a dash the derivative with respect

* Received by the editors July 19, 1984, and in final revised form November 16, 1984.
† Dipartimento di Matematica, Università di Torino, Via Carlo Alberto 10, 10123 Torino, Italy.

to $s \in J$. Let us consider a right-handed orthonormal reference frame $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ fixed in space. We assume that the rod is naturally straight with the end $s=0$ at the origin and its axis $\mathscr{C}$ along the direction $\mathbf{e}_3$. We take such a configuration as a *reference configuration*. We describe the *deformed configuration* by three vector functions $\mathbf{r}$, $\mathbf{a}_1$, $\mathbf{a}_2$ of the variable $s \in J$, such that $\mathbf{r}(s)$ is the *position vector* joining the origin $t=0$ and the centroid $t=s$, and $\mathbf{a}_1(s)$, $\mathbf{a}_2(s)$ are two orthonormal vectors fixed flat to the cross-section Σ through the centroid s. Define

$$\mathbf{a}_3(s) = \mathbf{a}_1(s) \times \mathbf{a}_2(s),$$

and assume that the rod can suffer neither extension nor shear; that is that (Kirchhoff hypotheses)

$$\mathbf{r}'(s) = \mathbf{a}_3(s). \tag{1}$$

Let $\mathbf{u}(s)$ be the unique vector satisfying

$$\mathbf{a}_j'(s) = \mathbf{u}(s) \times \mathbf{a}_j(s), \qquad j = 1, 2, 3.$$

Let

$$\mathbf{r}(s) = \sum_{i=1}^{3} x_i(s)\mathbf{e}_i$$

and

$$\mathbf{u}(s) = \sum_{i=1}^{3} u_i(s)\mathbf{a}_i(s).$$

In our theory u_j are the *strains*. From (1) we have

$$x_3'(s) = \left[1 - \left(x_1'^2(s) + x_2'^2(s)\right)\right]^{1/2}. \tag{2}$$

Let θ, ψ, ϕ be the (English) Euler angles describing the rotation of $\{\mathbf{a}_i(s)\}$ with respect to $\{\mathbf{e}_i\}$ (see [7, Article 253]), and let

$$\alpha = \frac{\pi}{2} - (\psi + \phi);$$

then it is easy to compute u_j in terms of x_1', x_2' and α, obtain (see [2, §1.1]):

$$\begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} = \begin{bmatrix} \cos\alpha & -\sin\alpha & 0 \\ \sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} x_1'' - \dfrac{x_1' x_3''}{1 + x_3'} \\ x_2'' - \dfrac{x_2' x_3''}{1 + x_3'} \\ -\alpha' + \dfrac{x_1'' x_2' - x_1' x_2''}{1 + x_3'} \end{bmatrix}. \tag{3}$$

Actually it is easy to realize that $\{\mathbf{a}_i(s)\}$ can be obtained from $\{\mathbf{e}_i\}$ by rotation $-\alpha$ around the $\mathbf{e}_3$-axis (torsion) followed by rotation around an axis in the $(\mathbf{e}_1, \mathbf{e}_2)$-plane and taking $\mathbf{e}_3$ onto $\mathbf{a}_3(s)$ (flexure). This makes it clearer that the state of the strain may be described purely in terms of x_1, x_2 and α.

2. Deformation energy. Following the usual director approach (see [1] and [2]), we assume that the *deformation energy* is given by

$$\int_0^1 \mathcal{W}(\mathbf{u}(s))\,ds, \tag{4}$$

where $\mathcal{W}$ is a smooth function such that $\mathcal{W}(0)=0$ and satisfying suitable symmetry hypotheses, which we describe below. We assumed that $\mathcal{W}$ does not depend explicitly on s; this means that the cross-sections do not depend on s, that is that the rod is *prismatic*. As we said above, $\mathcal{W}$ satisfies also suitable symmetry conditions which reflect the geometry of the cross-section. More precisely, the usual Kirchhoff transverse isotropy conditions state that if the principal moments of inertia of the cross-section are equal, then $\mathcal{W}$ is invariant with respect to the action of $\mathbb{O}(2)\oplus\mathbb{Z}_2\subset\mathbb{O}(3)$ on (u_1,u_2,u_3), defined by

$$(\gamma,\varepsilon)\cdot\mathbf{u}=\begin{bmatrix}\gamma_{11} & \gamma_{12} & 0\\ \gamma_{21} & \gamma_{22} & 0\\ 0 & 0 & \varepsilon\end{bmatrix}\cdot\begin{bmatrix}u_1\\ u_2\\ u_3\end{bmatrix},$$

where $\gamma=[\gamma_{jk}]$ is an orthogonal matrix and $\varepsilon=\pm 1$. The invariance of $\mathcal{W}$ with respect to the second summand, i.e. to $\mathbb{Z}_2$, represents indifference with respect to left-handed and right-handed twist. As regards the first summand, i.e. $\mathbb{O}(2)$, we already pointed out in [2, §1.2], that this hypothesis does not take completely into account the geometry of the cross-section Σ. For example, if Σ is a regular n-gon, its group of symmetries is $\mathbb{D}_n$ rather than $\mathbb{O}(2)$; thus it is more natural to *assume that $\mathcal{W}$ is invariant only under the action of $\Gamma_\Sigma\oplus\mathbb{Z}_2\subset\mathbb{O}(3)$, where Γ_Σ is the subgroup of $\mathbb{O}(2)$ which leaves invariant the cross-section Σ*. As we already said, if Σ is a regular n-gon we have $\Gamma_\Sigma=\mathbb{D}_n$; while if Σ is a rectangle, then $\Gamma_\Sigma=\mathbb{Z}_2\oplus\mathbb{Z}_2$.

Let us assume that the cross-section Σ is a rectangle with sides $2b_1$ and $2b_2$ as in Fig. 1. Define by $\delta=b_2-b_1$ the *aspect* of the cross-section. When $\delta=0$, Σ becomes a

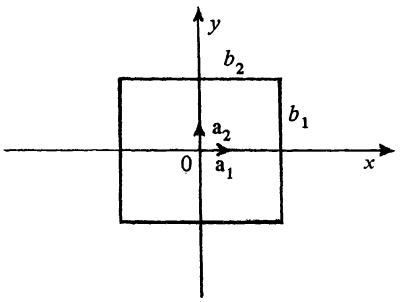


FIG. 1

square. On the ground of what we said above, *we assume that $\mathcal{W}$ is a smooth function of* $\mathbf{u}$ *and* δ, *invariant with respect to the action of $\mathbb{Z}_2\oplus\mathbb{Z}_2\oplus\mathbb{Z}_2$ when $\delta\neq 0$ and with respect to the action of $\mathbb{D}_4\oplus\mathbb{Z}_2$ when $\delta=0$*. Hence $\mathcal{W}$ is an invariant of the reflections $u_j\to -u_j$, $j=1,2,3$; thus, in a similar way as in [2, Prop. 1.13] one can prove that $\mathcal{W}$ depends on u_j^2 only. Therefore there exists a smooth function K such that

$$\mathcal{W}(\mathbf{u},\delta)=K\left(u_1^2,u_2^2,u_3^2,\delta\right). \tag{5}$$

Of course when $\delta=0$, that is the cross-section Σ is a square, K must also be an invariant of the rotation $\pi/2$ in the plane $u_3=0$; therefore by [2, Prop. 1.13] (with $n=4$) there exists a smooth function H such that

$$K(u_1^2,u_2^2,u_3^2,0)=H(\tau_1,\tau_2,\tau_3), \tag{6}$$

where $\tau_1=u_1^2+u_2^2$, $\tau_2=u_1^4-6u_1^2u_2^2+u_2^4$, $\tau_3=u_3^2$. In the following we shall write:

$$K(u^2,\delta)=K(u_1^2,u_2^2,u_3^2,\delta), \qquad H(\tau)=H(\tau_1,\tau_2,\tau_3).$$

Naturally if we expand K up to first order, we must obtain the classical quadratic energy of a rod (see [6, §18]):

$$K(u^2,\delta)=\frac{E_0}{2}\big(I_1(\delta)u_1^2+I_2(\delta)u_2^2\big)+\frac{C(\delta)}{2}u_3^2+\text{h.o.t.}, \tag{7}$$

where E_0 is the Young modulus, $I_1(\delta)$ and $I_2(\delta)$ are the principal moments of inertia of the cross-section and $C(\delta)$ is the torsional rigidity. It is possible to verify by a direct computation (a bit long for $C(\delta)$) that I_1, I_2 and C depend smoothly on δ, that is consistent with our assumptions on K. Moreover the following is true:

$$I_1(\delta)<I_2(\delta) \quad \text{if and only if} \quad \delta<0. \tag{8}$$

3. The variational problem. We can now give the *energy functional* f to be (locally) minimised by the rod. We hold fixed the end $s=0$ and apply a *terminal load force* P along the axis of the rod at the end $s=1$. Thus the *total energy* of the rod is the deformation energy minus the work done by P in moving along its line of action. The end $s=1$ moves ($x_3(0)=0$):

$$1-x_3(1)=\int_0^1\big(1-x_3'(s)\big)\,ds.$$

Thus from (4) and (5) we have

$$f(x_1,x_2,\alpha,P,\delta)=\int_0^1\big\{K(u^2(s),\delta)-P\big(1-x_3'(s)\big)\big\}\,ds. \tag{9}$$

We consider the rod clamped at both ends; that is satisfying the following set of boundary conditions:

$$x_1(0)=x_1(1)=x_2(0)=x_2(1)=0, \tag{10}$$

$$x_1'(0)=x_1'(1)=x_2'(0)=x_2'(1)=0, \tag{11}$$

$$\alpha(0)=\alpha(1)=0. \tag{12}$$

Conditions (11) and (12) mean that the ends $s=0$ and $s=1$ cannot rotate. In particular it is possible to prove that boundary conditions (10) to (12) correspond to have (see [2, §1.2])

$$\mathbf{a}_1(0)=\mathbf{a}_1(1)=\mathbf{e}_2, \quad \mathbf{a}_2(0)=\mathbf{a}_2(1)=-\mathbf{e}_1, \quad \mathbf{a}_3(0)=\mathbf{a}_3(1)=\mathbf{e}_3,$$

which fix the position of the rod in the unstressed state.

In order to state our variational problem, define the following Banach space

$$X=C_0^2(J)\times C_0^2(J)\times C_0^1(J),$$

where $C_0^k(J)=\{g:J\to\mathbb{R}\mid g$ is continuous with its derivatives $g^{(j)}$ for $0\leqq j\leqq k$ and $g^{(j)}(0)=g^{(j)}(1)=0$, for $0\leqq 2j\leqq k\}$. Moreover $C_0^k(J)$ is equipped with the maximum norm $|g_k|=\sup\{|g^{(j)}(s)|:s\in J$ and $0\leqq j\leqq k\}$ and X with the product-norm $\|x\|_X$. We denote by $x=(x_1,x_2,\alpha)$ the elements of X. Let

$$\Omega=\left\{(x,P,\delta)\in X\times\mathbb{R}\times\mathbb{R}\mid\sup_{s\in J}\left(x_1'^2(s)+x_2'^2(s)<1\right)\right\}.$$

It is easy to see that Ω is open and that by (2) and (3) we have $f\in C^\infty(\Omega,\mathbb{R})$.

The equilibria of the rod are given by the critical points with respect to $x\in X$ of the energy functional $f:\Omega\to\mathbb{R}$. In particular we are interested in (statically) *stable equilibria*, which correspond to (strict) *minima* of f. Let

$$P_j(\delta)=4\pi^2E_0I_j(\delta)=E_0I_j(\delta)\inf_{\phi\in C_0^2(J)}\frac{\int_0^1\phi''^2\,ds}{\int_0^1\phi'^2\,ds},\qquad j=1,2. \tag{13}$$

PROPOSITION 1. *For each $P\in\mathbb{R}$ and $\delta\in\mathbb{R}$, the unstressed configuration $(x_1,x_2,\alpha)=(0,0,0)$ is a critical point of the energy functional f. For $P<\inf\{P_1(\delta),P_2(\delta)\}$ the unstressed configuration $(0,0,0)$ corresponds to a strict (local) minimum of the energy functional f and hence to a stable equilibrium. Finally the second derivative $D_x^2f(0,P,\delta)$ is a degenerate quadratic form for $P=P_j(\delta)$, $j=1,2$.*

Proof. From (9), (5), (7), (3) and (2) one can compute easily the second derivative at $(0,P,\delta)$:

$$D_x^2f(0,P,\delta)[x,x]=\int_0^1\left\{E_0\left(I_1(\delta)x_1''^2+I_2(\delta)x_2''^2\right)+C(\delta)\alpha'^2-P\left(x_1'^2+x_2'^2\right)\right\}ds$$

and the result follows immediately. □

From Proposition 1 it follows immediately that $(0,P_j(\delta),\delta)$, $j=1,2$ are possible bifurcation points for f. On the other hand we have $P_1(0)=P_2(0)$, consequently it makes sense to state the following *perturbed variational bifurcation problem*: *find the number of (stable) critical points of f for (x,P,δ) near $(0,P_0,0)$, where $P_0=P_1(0)=P_2(0)$.*

4. Bifurcation analysis. In order to solve the problem stated above, we first reduce it to finite dimension, then we employ singularity theory. By applying Magnus' splitting lemma (see [2, Thm. 2.1] and [8]), we reduce the functional f to a function $\tilde f$ on a finite dimensional space in the same way as we did in [2]. Here we outline the reduction steps, referring to [2, Chap. 2] for the details. Let $I_0=I_1(0)=I_2(0)$; then $P=P_0$ $(=P_1(0)=P_2(0))$ is the first eigenvalue of the boundary-value problem

$$\begin{aligned}&E_0I_0\phi^{(4)}-P\phi''=0,\\&\phi(0)=\phi(0)=0,\\&\phi'(0)=\phi'(0)=0.\end{aligned}$$

The relevant eigenfunction is $\phi_0(s)=1-\cos 2\pi s$. Let

$$V=\left\{(\eta_1\phi_0(s),\eta_2\phi_0(s),0)\in X\mid\eta_1,\eta_2\in\mathbb{R}\right\}.$$

Define the Banach space

$$Y=C^0(J)\times C^0(J)\times C^0(J),$$

where $C^0(J)=\{g: J\to\mathbb{R} \mid g \text{ is continuous}\}$ and the continuous linear mapping $\omega: Y\to X^*$ (where X^* is the dual space to X) by

$$\langle y,x\rangle=\sum_{j=1}^{2}\int_0^1 y_j x_j''\,ds+\int_0^1 y_3\alpha'\,ds,$$

where $x=(x_1,x_2,\alpha)\in X$ and $y=(y_1,y_2,y_3)\in Y$ and $\langle\cdot,\cdot\rangle$ is the duality between X^* and X. Define $F:\Omega\to Y$ by

$$F(x,P,\delta)=\Bigg(\sum_{j=1}^{3}\left(K_{,u_j}u_{j,x_1''}-\int_0^1 K_{,u_j}u_{j,x_1'}\,dt\right)+P\int_0^1\left(1-x_3'\right)_{,x_1'}dt,$$

$$\sum_{j=1}^{3}\left(K_{,u_j}u_{j,x_2''}-\int_0^1 K_{,u_j}u_{j,x_2'}\,dt\right)+P\int_0^1\left(1-x_3'\right)_{,x_2'}dt,$$

$$\sum_{j=1}^{3}\left(K_{,u_j}u_{j,\alpha'}-\int_0^1 K_{,u_j}u_{j,\alpha}\,dt\right)\Bigg),$$

where a subscript following a comma indicates partial differentiation. We have $F\in C^\infty(\Omega, Y)$. Moreover, by an integration by parts one verifies that

$$D_x f(x,P,\delta)[\hat{x}]=\langle F(x,P,\delta),\,\hat{x}\rangle.$$

Let

$$T=D_x F(0,P_0,0).$$

Following the same steps as in [2, §2.3], we obtain that:

1) There exist decompositions $X=V\oplus Z$ and $Y=W\oplus Q$ such that $V=T^{-1}W$ and $T_{|Z}: Z\to Q$ is an isomorphism.

2) There exists a neighborhood U of $(0,P_0,0)\in X\times\mathbb{R}\times\mathbb{R}$ such that:

a) There exists a smooth mapping $h: U'\to Z$, with $U'=U\cap(V\times\mathbb{R}\times\mathbb{R})$, which is the unique solution $z=h(v,P,\delta)$ of the equation

$$P_Q F(v\oplus z,P,\delta)=0,$$

where P_Q is the orthogonal projection onto Q.

b) If we let $\eta=(\eta_1,\eta_2)\in\mathbb{R}^2$, $v=(\eta_1\phi_0,\eta_2\phi_0,0)\in V$, $\lambda=P-P_0$ and $\tilde{f}(\eta,\lambda,\delta)=f(v\oplus h(v,P,\delta),\,P,\delta)$, then

$$(\eta,\lambda,\delta)\to(v\oplus h(v,P,\delta),\,P,\delta)$$

defines a one-to-one and onto correspondence between the critial points of $\tilde{f}(\eta,\lambda,\delta)$ in U' and those of f in U. Moreover this correpondence preserves the minima.

3) $\tilde{f}(\eta,\lambda,\delta)$ is an invariant of the action of $\mathbb{Z}_2\oplus\mathbb{Z}_2$ on $\mathbb{R}^2$ generated by the symmetries $(\eta_1,\eta_2)\to(-\eta_1,\eta_2)$ and $(\eta_1,\eta_2)\to(\eta_1,-\eta_2)$.

In this way we reduce our problem to finding the number of (stable) critical points of $\tilde{f}$ for fixed λ and for (η,λ,δ) near $(0,0,0)$. This is equivalent to solving the following perturbed bifurcation problem:

$$G(\eta,\lambda,\delta)=0, \tag{14}$$

where $G = D_\eta \tilde{f}$, λ is the bifurcation parameter and δ is the perturbation parameter. Of course the minima of $\tilde{f}$, i.e., the (statically) stable equilibria, correspond to *positive* eigenvalues of the Jacobian matrix $D_\eta G$. Finally we have that the invariance of $\tilde{f}$ forces G to be $\mathbb{Z}_2 \oplus \mathbb{Z}_2$-equivariant i.e.:

$$G_1(-\eta_1, \eta_2, \lambda, \delta) = -G_1(\eta_1, \eta_2, \lambda, \delta),$$
$$G_2(-\eta_1, \eta_2, \lambda, \delta) = G_2(\eta_1, \eta_2, \lambda, \delta),$$

and

$$G_1(\eta_1, -\eta_2, \lambda, \delta) = G_1(\eta_1, \eta_2, \lambda, \delta),$$
$$G_2(\eta_1, -\eta_2, \lambda, \delta) = -G_2(\eta_1, \eta_2, \lambda, \delta),$$

where $G_j = \partial \tilde{f} / \partial \eta_j$, $j = 1, 2$.

In order to study the bifurcation problem (14), we employ, in the case of the symmetry group being $\mathbb{Z}_2 \oplus \mathbb{Z}_2$, the theory formulated in [4]. Recall briefly that two bifurcation problems G and $\hat{G}$ are $\mathbb{Z}_2 \oplus \mathbb{Z}_2$-equivalent if (see again [4, §§1 and 4])

$$\hat{G}(\eta, \lambda) = T(\eta, \lambda) G(E(\eta, \lambda), L(\lambda))$$

for some smooth family of invertible matrices T and a smooth diffeomorphism $(\eta, \lambda) \to (E(\eta, \lambda), L(\lambda))$ such that $(\partial L / \partial \lambda)(0) > 0$. Moreover we require that T and E be $\mathbb{Z}_2 \oplus \mathbb{Z}_2$-equivariant. It is easy to see that such an equivalence preserves the number of solutions for fixed λ, which is the information we are looking for.

Coming back to our problem, we have that in [2, Thm. 5.13] we have proven that if the function H in (6) is such that $(\partial H / \partial \tau_2)(0) \neq 0$ and the rod is prismatic, then $G(\eta, \lambda, 0)$ is $\mathbb{D}_4$-equivalent, hence $\mathbb{Z}_2 \oplus \mathbb{Z}_2$-equivalent, to

$$(A\eta_1^2 + \eta_2^2) - \lambda \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} + \varepsilon \begin{bmatrix} \eta_1^3 - 3\eta_1\eta_2^2 \\ \eta_2^3 - 3\eta_1^2\eta_2 \end{bmatrix} = 0, \tag{15}$$

where $A > 1$ and $\varepsilon = [\operatorname{sgn}(\partial H / \partial \tau_2)(0)]1$. Moreover the above equivalence preserves the signature of the real part of the eigenvalues of $D_\eta G$, that is the stability assignments of the solutions. This essentially solves our problem for $\delta = 0$.

It is worthwhile to remark that if H does not depend on τ_2 (and so in particular $(\partial H / \partial \tau_2)(0) = 0$) then one can prove that $g(\eta, \lambda, 0)$ has infinite $\mathbb{Z}_2 \oplus \mathbb{Z}_2$-codimension and our analysis fails. This is the main reason to consider nonlinear constitutive equations in our model.

Now we study $G(\eta, \lambda, \delta) = 0$ for δ near 0. Rewrite (15) as

$$\begin{bmatrix} (A+\varepsilon)\eta_1^3 + (A-3\varepsilon)\eta_1\eta_2^2 - \lambda\eta_1 \\ (A-3\varepsilon)\eta_1^2\eta_2 + (A+\varepsilon)\eta_2^3 - \lambda\eta_2 \end{bmatrix} = 0, \tag{16}$$

then multiply (16) by the matrix

$$\begin{bmatrix} \sqrt{A+\varepsilon} & 0 \\ 0 & \sqrt{A+\varepsilon} \end{bmatrix}$$

and divide η_j by $\sqrt{A+\varepsilon}$. We obtain that (16) is $\mathbb{Z}_2 \oplus \mathbb{Z}_2$-equivalent to

$$\begin{bmatrix} \eta_1^3 + \mu\eta_1\eta_2^2 - \lambda\eta_1 \\ \nu\eta_1^2\eta_2 + \eta_2^3 - \lambda\eta_2 \end{bmatrix} = 0, \tag{17}$$

with

$$\mu=\nu=\frac{A-3\varepsilon}{A+\varepsilon}. \tag{18}$$

Now (17) is a *double cusp bifurcation problem* and has been studied in [4, §4] and in [9, §6]. If it satisfies the nondegeneracy conditions

$$\mu\neq 1, \quad \nu\neq 1, \quad \mu\nu\neq 1, \tag{19}$$

it has universal unfolding

$$\begin{bmatrix} \eta_1^3+\tilde{\mu}\eta_1\eta_2^2-\lambda\eta_1 \\ \tilde{\nu}\eta_1^2\eta_2+\eta_2^3-(\lambda+a)\eta_2 \end{bmatrix}=0, \tag{20}$$

with $\tilde{\mu}$ and $\tilde{\nu}$ near μ and ν and a near 0. The Universal Unfolding Theorem (see [4, Thm. 1.8]) says that $G(\eta,\lambda,\delta)$ factors through (20), that is that there exists a smooth mapping $\delta\to(\tilde{\mu}(\delta),\tilde{\nu}(\delta),a(\delta))$ such that for each δ near 0, G and (20) are $\mathbb{Z}_2\oplus\mathbb{Z}_2$-equivalent. Now the nondegeneracy conditions (19) split the (μ,ν)-plane into 7 regions such that for fixed a all the bifurcation diagrams inside each region are topologically equivalent (see [9, §6]). Now if $(\tilde{\mu}(\delta),\ \tilde{\nu}(\delta))$ belongs to one of these regions for $\delta=0$, it will stay there also for each small $\delta\neq 0$. It follows that for a given δ our problem is associated to one of the regions according to the value of $\tilde{\mu}(0)$ and $\tilde{\nu}(0)$. Finally in each region there are three different diagrams according as $a(\delta)\lesseqgtr 0$ (see [9, §6]). Now for

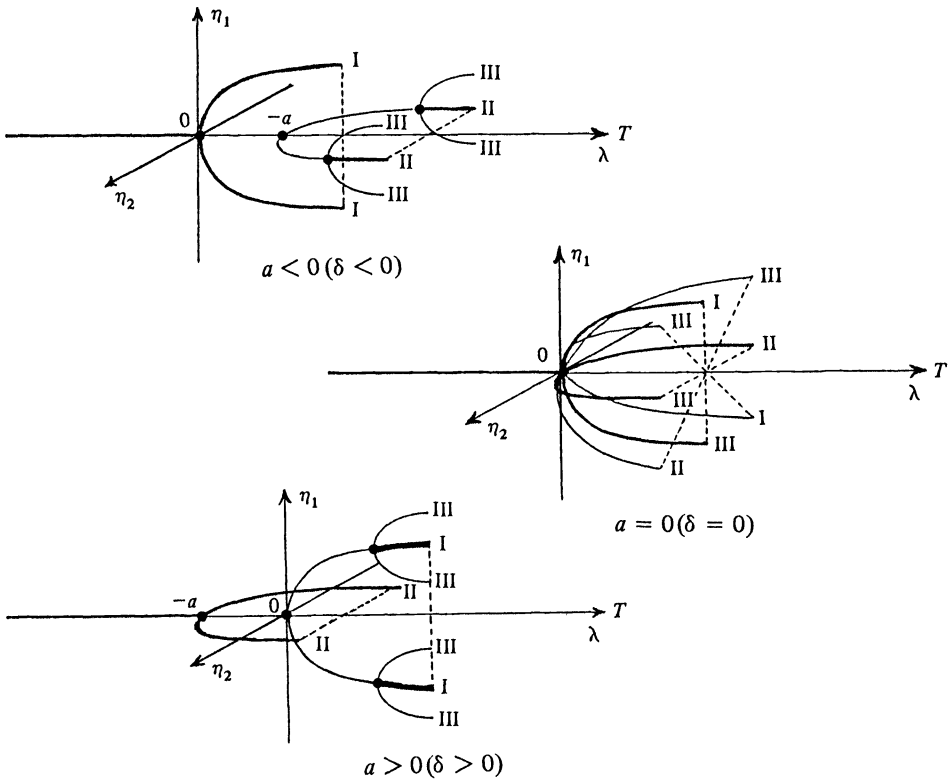


FIG. 2

our problem we have $\tilde{\mu}(0)=\tilde{\nu}(0)=(A-3\varepsilon)/(A+\varepsilon)$, thus there are only two possibilities: $(A-3\varepsilon)/(A+\varepsilon)\lesseqgtr 1$, corresponding to $\varepsilon=\pm 1$. So we have inspected all the possible cases and we are able to give the bifurcation diagrams. However, first we prefer to make some observations:

1) The bifurcation diagrams relate only to the number of solutions for fixed λ, their stability and their buckling features, as explained below.

2) As we did in [2, §§3.4 and 5.3], we can classify the solutions in three types, besides the trivial ones, according to their symmetry group and the kind of buckling the rod undergoes; see Table 1.

TABLE 1

Solution type	Symmetry group	Buckling
T	$\mathbb{Z}_2\oplus\mathbb{Z}_2$	trivial equilibrium
I	$\mathbb{Z}_2$ (reflection through η_2-axis)	in the plane $x_2=0$
II	$\mathbb{Z}_2$ (reflection through η_1-axis)	in the plane $x_1=0$
III	$\{1\}$	spatial

2) Table 1 allows us to relate the sign of a to the sign of δ. In fact $-a$ is the bifurcation point of solutions of type II and this is negative if and only if the rod buckles first in the plane $x_1=0$. Now by Proposition 1 this happens if and only if $P_2(\delta)<P_1(\delta)$, that is if and only if $\delta>0$, by (8) and (13).

5. Bifurcation diagrams. In the diagrams in Figs. 2 and 3 we draw stable equilibria (i.e. minima) in heavy black.

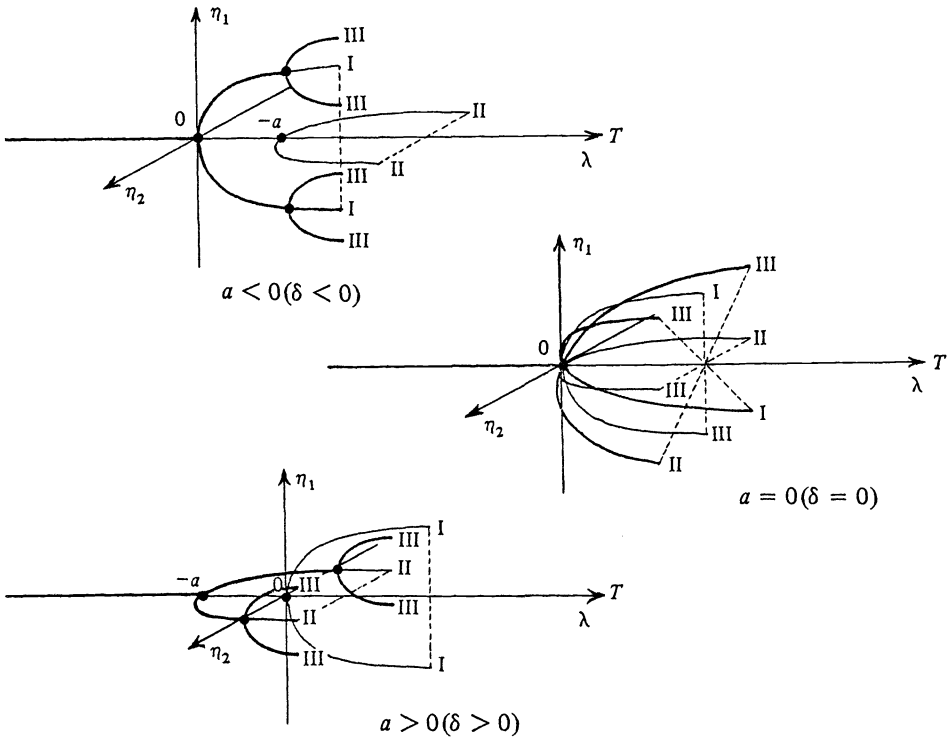


FIG. 3

REFERENCES

[1] S. S. ANTMAN AND C. S. KENNEY, *Large buckled states of non-linearly elastic rods under torsion, thrust and gravity*, Arch. Rat. Mech. Anal., 76 (1981), pp. 339–354.

[2] E. BUZANO, G. GEYMONAT AND T. POSTON, *Post-buckling behavior of a non-linearly hyperelastic thin rod with cross-section invariant under the dihedral group D_n*, in Arch. Rat. Mech. Anal., to appear.

[3] E. BUZANO AND A. RUSSO, *Non-axisymmetric post-buckling behavior of a complete thin cylindrical shell under axial compression*, to appear.

[4] M. GOLUBITSKY AND D. SCHAEFFER, *Imperfect bifurcation in the presence of symmetry*, Comm. Pure Appl. Math., 32 (1979), pp. 21–98.

[5] K. KOVARI, *Räumliche Verzweigungsprobleme des dünnen elastischen Stabes mit endlichen Verformungen*, Ing. Arch., 37 (1969), pp. 393–416.

[6] L. LANDAU AND E. M. LIFSCHITZ, *Theory of Elasticity*, Pergamon Press, New York, 1959.

[7] A. E. LOVE, *A Treatise on the Mathematical Theory of Elasticity*, Cambridge Univ. Press, Cambridge, 1927; (Dover reprint, New York, 1944).

[8] R. MAGNUS, *A splitting lemma for non-reflexive Banach spaces*, Math. Scan., 46 (1980), pp. 118–128.

[9] D. SCHAEFFER AND M. GOLUBITSKY, *Boundary conditions and mode-jumping in the buckling of a rectangular plate*, Comm. Math. Phys., 69 (1979), pp. 209–236.

[10] D. SCHAEFFER AND M. GOLUBITSKY, *Bifurcation analysis near a double eigenvalue of a model chemical reaction*, Arch. Rat. Mech. Anal., 75 (1981), pp. 315–347.

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

SIAM J. MATH. ANAL.
Vol. 17, No. 2, March 1986

004

EXACT BOUNDARY CONDITIONS AT AN ARTIFICIAL BOUNDARY FOR PARTIAL DIFFERENTIAL EQUATIONS IN CYLINDERS*

THOMAS HAGSTROM† AND H. B. KELLER‡

Abstract. The numerical solution of partial differential equations in unbounded domains requires a finite computational domain. Often one obtains a finite domain by introducing an artificial boundary and imposing boundary conditions there. This paper derives exact boundary conditions at an artificial boundary for partial differential equations in cylinders. An abstract theory is developed to analyze the general linear problem. Solvability requirements and estimates of the solution of the resulting finite problem are obtained by use of the notions of exponential and ordinary dichotomies. Useful representations of the boundary conditions are derived using separation of variables for problems with constant tails. The constant tail results are extended to problems whose coefficients obtain limits at infinity by use of an abstract perturbation theory. The perturbation theory approach is also applied to a class of nonlinear problems. General asymptotic formulas for the boundary conditions are derived and displayed in detail.

AMS(MOS) subject classifications. Primary 35A05, 35A40, 35C20, 65N99

Key words. artificial boundary conditions, asymptotic expansions for PDE's

1. Introduction. Many of the boundary value problems arising in applied mathematics are given on unbounded domains. Examples include the problems of fluid flow and wave propagation in channels or past bodies. The numerical solution of these problems, however, requires a finite domain. In this paper, we develop a theory for the exact reduction of a boundary value problem for a partial differential equation on an unbounded cylindrical domain to a problem on a bounded domain. That is, an "artifical" boundary is introduced and the proper boundary condition to be imposed there is derived. In other works, [8] and [9], we use our theory to solve nonlinear problems of both elliptic and parabolic type.

For ordinary differential equations, exact reduction theories have been developed by many authors: de Hoog and Weiss [5], Keller and Lentini [11], Jepson and Keller [10] and Markowich [12]. Few works on artificial boundary conditions for partial differential equations, on the other hand, have discussed exact conditions. An exception is the paper of Gustafsson and Kreiss [6], where the form of the proper conditions for a general hyperbolic problem is derived. They go on to find representations of the exact conditions in various simple cases for problems of both hyperbolic and elliptic type.

We illustrate the derivation of exact conditions with the following example:

$$
\begin{aligned}
&\text{a)} \quad \nabla^2 u + a(x,\mathbf{y})u = f(x,\mathbf{y}), \quad (x,\mathbf{y}) \in [0,\infty)\times\Omega, \quad \Omega \subset R^{n-1},\\
&\text{b)} \quad c(x,\mathbf{y})\frac{\partial u}{\partial \nu}(x,\mathbf{y}) + d(x,\mathbf{y})u(x,\mathbf{y}) = \gamma_\Omega(x,\mathbf{y}), \qquad \mathbf{y} \in \partial\Omega,\\
&\text{c)} \quad \alpha(\mathbf{y})\frac{\partial u}{\partial x}(0,\mathbf{y}) + b(\mathbf{y})u(0,\mathbf{y}) = \gamma_0(\mathbf{y}), \qquad \mathbf{y} \in \Omega,\\
&\text{d)} \quad \lim_{x\to\infty} u(x,\mathbf{y}) = 0,\\
&\text{e)} \quad c(x,\mathbf{y}) = c_\infty(\mathbf{y}), \quad d(x,\mathbf{y}) = d_\infty(\mathbf{y}), \quad a(x,\mathbf{y}) = a_\infty(\mathbf{y}), \quad x \geqq x_0,\\
&\qquad f(x,\mathbf{y}) = \gamma_\Omega(x,\mathbf{y}) = 0, \qquad x \geqq x_0.
\end{aligned} \tag{1.1}
$$

*Received by the editors April 5, 1984. This research was sponsored by the U. S. Army under contract DAAG29-80-C-0041, and supported in part by the U. S. Department of Energy under contract DE-AS03-76SF-00767.

†Department of Applied Mathematics and Statistics, State University of New York, Stony Brook, New York 11794.

‡Applied Mathematics, California Institute of Technology, Pasadena, California 91125.

(We say that (1.1) has a constant tail, due to condition (1.1e).) We choose some point $x=\tau \geqq x_0$ as the location of the artificial boundary. In the "tail", $x \geqq \tau$, we have:

$$
\begin{aligned}
&\text{a)} \quad \nabla^2 u + a_\infty(\mathbf{y})u = 0, \qquad (x,\mathbf{y}) \in [\tau,\infty)\times\Omega; \\
&\text{b)} \quad c_\infty(\mathbf{y})\frac{\partial u}{\partial \nu}(x,\mathbf{y}) + d_\infty(\mathbf{y})u(x,\mathbf{y}) = 0, \qquad \mathbf{y}\in\partial\Omega; \\
&\text{c)} \quad \lim_{x\to\infty} u(x,\mathbf{y}) = 0.
\end{aligned} \tag{1.2}
$$

Problem (1.2) can be easily analyzed by separation of variables. Consider the following eigenvalue problem:

$$
\begin{aligned}
&\text{a)} \quad \nabla_{\mathbf{y}}^2 Y_n(\mathbf{y}) + a_\infty(\mathbf{y})Y_n(\mathbf{y}) = \omega_n Y_n(\mathbf{y}), \qquad \mathbf{y}\in\Omega, \\
&\text{b)} \quad c_\infty(\mathbf{y})\frac{\partial}{\partial \nu}Y_n(\mathbf{y}) + d_\infty(\mathbf{y})Y_n(\mathbf{y}) = 0, \qquad \mathbf{y}\in\partial\Omega; \\
&\text{c)} \quad \int_\Omega d\mathbf{y}\, Y_n^2(\mathbf{y}) = 1.
\end{aligned} \tag{1.3}
$$

Given certain assumptions on the boundary condition, (1.3b), the set of eigenfunctions, $\{Y_n\}$, is complete in that subspace of $L_2(\Omega)$ consisting of functions satisfying it. (See Berezanskii [3].) For simplicity, we further assume that the ω_n are distinct and that $\omega_n = 0$ is not an eigenvalue. We rewrite the $\{\omega_n\}$ in the following way:

$$
\begin{aligned}
&\omega_n = \alpha_n^2 > 0, \qquad n = 1,\cdots,m, \\
&\omega_n = -\lambda_n^2 < 0, \qquad n = m+1,\ m+2,\cdots.
\end{aligned} \tag{1.4}
$$

Expanding u in terms of the Y_n's,

$$
u(x,\mathbf{y}) = \sum_{n=1}^{\infty} c_n(x)Y_n(\mathbf{y}) \tag{1.5}
$$

problem (1.2) becomes:

$$
\begin{aligned}
&\text{a)} \quad c_n'' = \begin{cases} -\alpha_n^2 c_n, & n = 1,\cdots,m, \\ \lambda_n^2 c_n, & n = m+1,\ m+2,\cdots, \end{cases} \qquad x\in[\tau,\infty); \\
&\text{b)} \quad \lim_{x\to\infty} c_n(x) = 0, \qquad n = 1,2,\cdots.
\end{aligned} \tag{1.6}
$$

As (1.6a) can be trivially solved, we see that (1.6b) is satisfied if and only if:

$$
\begin{aligned}
&\text{a)} \quad c_n(\tau) = c_n'(\tau) = 0, \qquad n = 1,\cdots,m; \\
&\text{b)} \quad c_n'(\tau) = -\lambda_n c_n(\tau), \qquad n = m+1,\ m+2,\cdots.
\end{aligned} \tag{1.7}
$$

This allows us to replace (1.1) by an equivalent finite domain problem:

$$
\begin{aligned}
&\text{a)} \quad \nabla^2 u + a(x,\mathbf{y})u = f(x,\mathbf{y}), \quad (x,\mathbf{y})\in[0,\tau]\times\Omega; \\
&\text{b)} \quad c(x,\mathbf{y})\frac{\partial u}{\partial \nu}(x,\mathbf{y}) + d(x,\mathbf{y})u(x,\mathbf{y}) = \gamma_\Omega(x,\mathbf{y}), \quad \mathbf{y}\in\partial\Omega; \\
&\text{c)} \quad \alpha(\mathbf{y})\frac{\partial u}{\partial x}(0,\mathbf{y}) + b(\mathbf{y})u(0,\mathbf{y}) = \gamma_0(\mathbf{y}), \quad \mathbf{y}\in\Omega; \\
&\text{d)} \quad \int_\Omega d\mathbf{y}\, u(\tau,\mathbf{y})Y_n(\mathbf{y}) = \int_\Omega d\mathbf{y}\frac{\partial u}{\partial x}(\tau,\mathbf{y})Y_n(\mathbf{y}) = 0, \quad n = 1,\cdots,m; \\
&\qquad \int_\Omega d\mathbf{y}\frac{\partial u}{\partial x}(\tau,\mathbf{y})Y_n(\mathbf{y}) = -\lambda_n\int_\Omega d\mathbf{y}\, u(\tau,\mathbf{y})Y_n(\mathbf{y}), \quad n = m+1,\ m+2,\cdots.
\end{aligned} \tag{1.8}
$$

That is, (1.1) has a solution if and only if (1.8) does and the solutions agree on a finite domain.

In §2 of this work we derive boundary conditions for the reduction of a general partial differential equation in a semi-infinite cylindrical domain to a finite one. These turn out to be the requirement that the appropriate data at the artificial boundary lie in a certain affine set. We find it convenient to rewrite the problem as an ordinary differential equation in a Banach space, making transparent the connection between our reduction and the reduction theorems for the case of ordinary differential equations. In §3 we introduce the notion of a dichotomy for our abstract equation and use it to develop error estimates and solvability requirements for the finite problem.

We first consider the problem of representing the boundary conditions in §4. Here separation of variables is used to analyze constant tail problems such as the one presented above. The exact representation we obtain is equivalent to (1.8d) in that case.

We develop a perturbation theory to analyze nonconstant tail problems in §5. Assuming the limiting problem at infinity can be solved by separation of variables, a perturbation expansion of the exact boundary condition can be calculated. We carry out this expansion for the Helmholtz equation exterior to a body, recovering the conditions of Bayliss, Gunzburger and Turkel [2]. Finally, in §6, nonlinear problems are considered. We use the perturbation theory of the preceding section to prove, under certain conditions, the existence of an exact nonlinear boundary condition and to calculate an expansion which approximates it.

We note that many authors have derived boundary conditions for specific problems. We do not, in general, attempt to examine the connection between their conditions and ours. For more discussion of these connections as well as for a more extensive bibliography, the reader is referred to Hagstrom [7].

2. Basic linear reduction theorem. We consider abstract boundary value problems of the form:

$$
\begin{aligned}
&\text{a)} \quad \frac{du}{dx} = A(x)u + f(x), \quad 0 < x < \infty; \\
(2.1)\quad &\text{b)} \quad B_0 u(0) = \gamma_0, \\
&\text{c)} \quad \lim_{x \to \infty} B_\infty u(x) = 0.
\end{aligned}
$$

In addition we may impose:

$$
\text{d)} \quad \| u(x) \| \text{ bounded as } x \to \infty.
$$

For some Banach space, $\mathscr{B}$, we seek $u(x) \in \mathscr{B}$ for $x \in [0, \infty]$. We suppose that $A(x)$, B_0 and B_∞ are linear operators with domain in $\mathscr{B}$, to which we also constrain the range of $A(x)$. Finally, $f(x) \in \mathscr{B}$.

Problems of form (1.1) can be obtained from general partial differential equation problems in cylindrical domains. Specifically we consider

$$
(2.2) \qquad \left(\sum_{j=1}^{n} P_j\left(\mathbf{y}, \mathbf{x}, \frac{\partial}{\partial \mathbf{y}}\right) \frac{\partial^j}{\partial x^j} \right) \omega = g(x, \mathbf{y});
$$

on the cylindrical domain

$$
(x, \mathbf{y}) \in [0, \infty) \times \Omega, \qquad \Omega \subset \mathbf{R}^n.
$$

Homogeneous boundary conditions are imposed on $\partial\Omega$ involving ω and its normal derivatives;

$$\sum_{j=0}^{m} B_{\Omega,hj}(\mathbf{y})\frac{\partial^j \omega}{\partial \nu^j}(x,\mathbf{y})=0, \qquad \mathbf{y}\in\partial\Omega. \tag{2.3}$$

We further suppose that, subject to these boundary conditions,

$$P_n^{-1}\left(x,\mathbf{y},\frac{\partial}{\partial \mathbf{y}}\right)$$

exists for all x. Now (2.2) can be rewritten in the form of (2.1a) by introducing

$$u=\begin{pmatrix} \dfrac{\partial^{n-1}\omega}{\partial x^{n-1}} \\ \dfrac{\partial^{n-2}\omega}{\partial x^{n-2}} \\ \vdots \\ \omega \end{pmatrix}. \tag{2.4}$$

The space, $\mathscr{B}$, is some space of n-tuples of functions on Ω which satisfy the homogeneous boundary conditions, (2.3). It is necessary to eliminate inhomogeneous conditions on $\partial\Omega$ in order to reduce the problem to the abstract form. This can be accomplished by subtracting a function that satisfies the inhomogeneous condition. We note that the functions $B_{\Omega,j}$ affect P_n^{-1} and, ultimately, $A(x)$.

Returning to (2.1) we choose some finite point, $x=\tau$, and attempt to reduce the infinite problem on $[0,\infty)$ to a finite one on $[0,\tau]$. We define $\mathbf{A}(\tau;\mathrm{f})$, the admissible set of Cauchy data at $x=\tau$, as the set leading to solutions, u, in the tail, $x\in[\tau,\infty)$. More precisely we have:

DEFINITION 2.5. The set $\mathbf{A}(\tau;f)\subset\mathscr{B}$, the admissible set at $x=\tau$, is the set of all $u_0\in\mathscr{B}$ such that there exists $u(x)\in\mathscr{B}$, $x\in[\tau,\infty)$, satisfying:

$$\begin{aligned} &\text{a)} \quad \frac{du}{dx}=A(x)u+f(x), \qquad \tau<x<\infty, \\ &\text{b)} \quad u(\tau)=u_0; \end{aligned} \tag{2.5}$$

as well as (2.1c, d) as appropriate.

It is now possible to write down an exact reduction of (2.1) to a problem on a finite domain. We state the reduction as a theorem whose proof follows immediately from the definition of $\mathbf{A}(\tau;f)$.

THEOREM 2.6. *Problem* (2.1) *has a solution if and only if the following problem has a solution*:

$$\begin{aligned} &\text{a)} \quad \frac{d\omega}{dx}=A(x)\omega+f(x), \qquad x\in[0,\tau]; \\ &\text{b)} \quad B_0\omega(0)=\gamma_0; \\ &\text{c)} \quad \omega(\tau)\in\mathbf{A}(\tau;f). \end{aligned} \tag{2.6}$$

Furthermore, whenever (2.1) has a solution $u(x)$, (2.6) has a solution which is identical to u on $[0,\tau]$.

Proof. Suppose (2.6) has a solution. Then, by the definition of $\mathbf{A}(\tau;f)$, there exists $u^+(x)$, $x\in[\tau,\infty)$, satisfying (2.1a) and (2.1c, d) as appropriate as well as $\omega(\tau)=u^+(\tau)$. Define

$$u(x)=\begin{cases}\omega(x), & x\in[0,\tau],\\ u^+(x), & x\in[\tau,\infty).\end{cases}$$

Then, u is a solution of (2.1). Now, suppose that (2.1) has a solution. The restriction of u to $[\tau,\infty)$ satisfies (2.5) and (2.1c, d) and, hence, by the definition of $\mathbf{A}(\tau;f)$, $u(\tau)\in\mathbf{A}(\tau;f)$. This implies that the restriction of u to $[0,\tau]$ satisfies (2.6), completing the proof.

The set $\mathbf{A}(\tau;f)$ is an affine subset of $\mathscr{B}$. A convenient representation of $\mathbf{A}$ can be found in terms of its underlying linear subspace and some particular element of $\mathscr{B}$. We consider the homogeneous problem in the tail associated with (2.1):

$$\text{(2.7)}\quad \begin{aligned}&\text{a)}\quad \frac{dv}{dx}=A(x)v, \qquad x\in[\tau,\infty);\\ &\text{b)}\quad \lim_{x\to\infty} B_\infty v(x)=0,\end{aligned}$$

and, if (2.1d) is imposed,

$$\text{c)}\quad \|v(x)\| \text{ bounded as } x\to\infty.$$

We define $\mathscr{A}(\tau)$, the admissible space at $x=\tau$, as the set of all Cauchy data leading to solutions of (2.7). That is:

DEFINITION 2.8. The set $\mathscr{A}(\tau)\subset\mathscr{B}$, the admissible space at $x=\infty$, is the set of all $v_0\in\mathscr{B}$ such that (2.7) has a solution satisfying:

$$\text{(2.8)}\qquad v(\tau)=v_0.$$

We note that $\mathscr{A}(\tau)$ is independent of the inhomogeneous term in (2.1). We further require a particular solution, $u_p(x)$, which satisfies:

$$\text{(2.9)}\quad \begin{aligned}&\text{a)}\quad \frac{du_p}{dx}=A(x)u_p+f(x), \qquad x\in[\tau,\infty);\\ &\text{b)}\quad \lim_{x\to\infty} B_\infty u_p(x)=0;\end{aligned}$$

and, if (2.1d) is imposed

$$\text{c)}\quad \|u_p(x)\| \text{ bounded as } x\to\infty.$$

We note that if $\mathbf{A}(\tau;f)$ is nonempty, at least one such $u_p(x)$ must exist. It is now possible to prove:

THEOREM 2.10. *Let $u_0\in\mathscr{B}$. Then $u_0\in\mathbf{A}(\tau;f)$ if and only if, for any particular solution $u_p(x)$*

$$\text{(2.10)}\qquad u_0-u_p(\tau)\in\mathscr{A}(\tau).$$

Proof. The proof is an immediate consequence of the definitions of $\mathbf{A}(\tau;f)$, $\mathscr{A}(\tau)$ and $u_p(x)$ combined with the linearity of (2.1).

If we assume that there exists a projection operator, $Q(\tau)$, into $\mathscr{A}(\tau)$, we can rewrite (2.10):

$$(I-Q(\tau))(u_0-u_p(\tau))=0.$$

In particular, the boundary condition, (2.6c), can be replaced by:

$$(2.6c') \qquad (I-Q(\tau))\omega(\tau)=(I-Q(\tau))u_p(\tau).$$

We emphasize that $u_p(x)$ can be any particular solution.

Finally, we write down a corollary of Theorem 2.6 which concerns the uniqueness of solutions.

COROLLARY. *Suppose that for all* $v_0\in\mathscr{A}(\tau)$ *solutions to the Cauchy problem defined by* (2.7a) *and* (2.8) *are unique. Then* (2.6) *has a unique solution if and only if* (2.1) *does.*

Proof. Assuming uniqueness of solutions to (2.1) immediately yields uniqueness for (2.6). In the other direction, note that the assumption above guarantees the uniqueness of $u_+(x)$ which, combined with the uniqueness of the finite interval solution, implies the uniqueness of u.

3. Solvability of the finite problem. In this section we assume that solutions to the homogeneous Cauchy problems:

$$(3.1) \quad \text{a)} \quad \frac{dv}{dx}=A(x)v, \quad x_0\leqq x\leqq x_1 \text{ if } x_1>x_0; \quad x_1\leqq x\leqq x_0 \text{ if } x_0>x_1;$$
$$\text{b)} \quad v(x_0)=v_0;$$

are unique for all x_0, $x_1\in[0,\infty)$. We define a solution operator $S(x_1,x_0;A)$ in the following way:

DEFINITION 3.2. Let $v_0\in\mathscr{B}$. If there exists a solution, $v(x)$ to problem (3.1) then

$$(3.2) \qquad S(x_1,x_0;A)v_0=v(x_1).$$

Otherwise, v_0 is said to be outside the domain of $S(x_1,x_0;A)$.

The linearity of the differential equation implies the linearity of s. The stated uniqueness of solutions implies the consistency of the definition. Note that it is certainly necessary to restrict the domain of S for ill-posed Cauchy problems such as those which arise in the study of elliptic equations. Whenever S exists, however, it does have the familiar semi-group properties:

$$(3.3) \quad \text{a)} \quad S(x_1,x^*;A)S(x^*,x_0;A)=S(x_1,x_0;A),$$
$$\text{b)} \quad S(x_0,x_0;A)=I.$$

The notion of dichotomies is very useful in what follows. First we present definitions of exponential and ordinary dichotomies. These are adapted from Daletskiy and Krein [4], with some modifications required by the poosible nonexistence of solutions.

DEFINITION 3.4. We say that the problem

$$(3.4) \qquad \frac{dv}{dx}=A(x)v, \qquad x\in[0,\infty);$$

has an exponential dichotomy if, for any $x^* \in [0, \infty)$, the space $\mathscr{B}$ can be decomposed into a direct sum of subspaces $\mathscr{B}_-(x^*)$ and $\mathscr{B}_+(x^*)$ such that:

(3.5)
a) If $v \in \mathscr{B}_-(x^*)$ then, for some N_- and $\alpha_- > 0$
i) $S(x, x^*; A)\, v$ exists for any $x \geqq x^*$;
ii) $\| S(x, x^*; A) v \| \leqq N_- e^{-\alpha_-(x - x^*)} \| v \|$.

(3.6)
b) If $v \in \mathscr{B}_+(x^*)$ then, for some N_+ and $\alpha_+ > 0$
i) $S(x, x^*; A) v$ exists for any $x \leqq x^*$;
ii) $\| S(x, x^*; A) v \| \leqq N_+ e^{-\alpha_+(x^* - x)} \| v \|$.

(3.7)
c) There exists $\gamma > 0$, independent of x^*, such that

$$\inf_{\substack{u_\pm \in \mathscr{B}_\pm(x^*) \\ \|u_\pm\| = 1}} \| u_+ + u_- \| \geqq \gamma.$$

(This infimum is typically called the angular distance between $\mathscr{B}_+(x^*)$ and $\mathscr{B}_-(x^*)$.)

An ordinary dichotomy is defined as above except that $\alpha_\pm = 0$ is allowed. No "continuity" of the spaces as functions of x^* has so far been required. In general, we impose a sort of continuity in the form of the following "no-mixing" condition.

DEFINITION 3.8. The dichotomy (3.5–3.7) satisfies the no-mixing condition if whenever

(3.8)
a) $Q(x)$ is the projection operator into $\mathscr{B}_-(x)$,
b) $S(x_1, x_0; A)\, v$ exists

then

c) $Q(x_1) S(x_1, x_0; A) v = S(x_1, x_0; A) Q(x_0) v$.

Assuming that the homogeneous problem has a dichotomy in the tail and that $\mathscr{B}_-(x)$ coincides with the admissible space, $\mathscr{A}(x)$, it is possible to write down an integral expression for a particular solution, $u_p(x)$ which is valid whenever $\|f(x)\|$ is integrable;

$$u_p(x) = \int_\tau^x S(x, p; A) Q(p) f(p)\, dp - \int_x^\infty S(x, p; A)(I - Q(p)) f(p)\, dp. \tag{3.9}$$

(The validity of (2.9a) follows from the direct differentiation of (3.9) while (2.9b) is insured by the identity of $\mathscr{B}_-(x)$ and $\mathscr{A}(x)$ combined with the absolute convergence of the integrals.) Note that it is always the case that $\mathscr{B}_-(x) = \mathscr{A}(x)$ if there is an exponential dichotomy. Then, only boundedness of $\|f\|$ need be assumed.

Formula (3.9) is extremely useful in the development of a perturbation theory. For now, we simply use it to write down a new expression for the boundary condition, (2.6c):

$$(I - Q(\tau))\, \omega(\tau) = -\int_\tau^\infty S(\tau, p)(I - Q(p)) f(p)\, dp. \tag{3.10}$$

Extending the dichotomy to the entire interval, we now can prove an existence theorem for the finite boundary value problem (2.6).

THEOREM 3.11. *Suppose that solutions to all Cauchy problems* (3.1) *are unique for* x_0, $x_1 \in [0,\tau]$ *and that* (3.1a) *has a nonmixing ordinary dichotomy on* $[0,\tau]$ *with projector* $Q(x)$ *into* $\mathscr{B}_-(x)$. *Also assume that* $\mathscr{B}_-(\tau)=\mathscr{A}(\tau)$. *Then* (2.6) *has a solution for arbitrary* $f(x)$, $u_p(\tau)$ *and* γ_0 *in the range of* B_0 *if and only if the operator*

$$\Phi\omega = \begin{Bmatrix} (I-Q(0))\omega \\ B_0\omega \end{Bmatrix} \tag{3.11}$$

has an inverse with domain containing all pairs of the form:

$$\begin{pmatrix} 0 \\ \gamma \end{pmatrix}, \quad \gamma \in \text{Range}\,(B_0). \tag{3.12}$$

The solution is unique and bounded in terms of the inhomogeneous data if and only if this (*restricted*) *inverse is.*

Proof. We use the ordinary dichotomy defined by $Q(x)$ to solve certain initial value problems. Let

$$\omega_+(x) = S(x,\tau;A)(I-Q(\tau))u_p(\tau) + \int_\tau^x S(x,p;A)(I-Q(p))f(p)\,dp. \tag{3.13}$$

This exists for all x on $[0,\tau]$ by the definition of Q. If we seek solutions to (2.6) in the form

$$\omega(x) = \omega_+(x) + \omega_-(x) \tag{3.14}$$

then ω is a solution if and only if ω_- solves

$$\begin{aligned} &\text{a)} \quad \frac{d\omega_-}{dx} = A(x)\omega_- + Q(x)f(x); \\ &\text{b)} \quad B_0\omega_-(0) = \gamma_0 - B_0\omega_+(0); \\ &\text{c)} \quad (I-Q(\tau))\omega_-(\tau) = 0. \end{aligned} \tag{3.15}$$

We write $\omega_-(x)$ in the form:

$$\omega_-(x) = S(x,0;A)\omega_-(0) + \int_0^x S(x,p;A)Q(p)f(p)\,dp. \tag{3.16}$$

The integral term again exists by the definition of Q so that this representation is valid for any solution of (3.15a). By (3.15c) and (3.8c) we have:

$$0 = (I-Q(\tau))\omega_-(\tau) = S(\tau,0;A)(I-Q(0))\omega_-(0);$$

which, by the uniqueness of solutions to the Cauchy problem, implies

$$(I-Q(0))\omega_-(0) = 0.$$

Hence, we can find a solution to (3.15) if and only if we can simultaneously solve:

$$\begin{aligned} &(I-Q(0))\omega_-(0) = 0; \\ &B_0\omega_-(0) = \gamma_0 - B_0\omega_+(0); \end{aligned}$$

which in component form yields (3.11), completing the proof.

Estimates of the solution in terms of the inhomogeneous data are now obtained from the explicit representation in terms of ω_+ and ω_-. Assume that

$$
\begin{aligned}
&\text{a)} \quad \|S(x,p;A)Q(p)\| \leqq K_-(x,p), \qquad 0 \leqq p \leqq x \leqq \tau;\\
&\text{b)} \quad \|S(x,p;A)(I-Q(p))\| \leqq K_+(x,p), \qquad 0 \leqq x \leqq p \leqq \tau;\\
&\text{c)} \quad \left\|\Phi^{-1}\begin{pmatrix}0\\ \gamma\end{pmatrix}\right\| \leqq K_\phi \|\gamma\|, \qquad \gamma \in \text{Range}(B_0);\\
&\text{d)} \quad \|B_0\| \leqq K_0.
\end{aligned} \tag{3.17}
$$

Then we have, directly estimating (3.13) and (3.16) and using the fact that $Q(0)\omega_-(0) = \omega_-(0)$,

$$
\begin{aligned}
\|\omega(x)\| \leqq K_-(x,0)\|\gamma_0\| &+ \max_{x\in[0,\tau]} \|f(x)\| \int_0^x K_-(x,p)\,dp\\
&+ \max_{x\in[0,\tau]} \|f(x)\| \left(\int_x^\tau K_+(x,p)\,dp + K_-(x,0)K_\phi K_0 \int_0^\tau K_+(0,p)\,dp \right)\\
&+ \|u_p(\tau)\| \left(K_+(x,\tau) + K_-(x,0)K_\phi K_0 K_+(0,\tau) \right).
\end{aligned} \tag{3.18}
$$

Equation (3.18) allows us to estimate the errors caused by approximations to $Q(\tau)$ and $u_p(\tau)$. Suppose we solve the following finite problem instead of (2.6):

$$
\begin{aligned}
&\text{a)} \quad \frac{d\omega_a}{dx} = A(x)\omega_a + f(x), \qquad 0 \leqq x \leqq \tau,\\
&\text{b)} \quad B_0\omega_a(0) = \gamma_0;\\
&\text{c)} \quad (I-Q^*(\tau))\omega_a(\tau) = (I-Q^*(\tau))u_p^*(\tau);
\end{aligned} \tag{3.19}
$$

where $Q^*(\tau)$ and $u_p^*(\tau)$ differ from $Q(\tau)$ and $u_p(\tau)$. We define the error, $e(x)$, by

$$e(x) \equiv \omega(x) - \omega_a(x)$$

and find that it satisfies:

$$
\begin{aligned}
&\text{a)} \quad \frac{de}{dx} = A(x)e, \qquad 0 \leqq x \leqq \tau,\\
&\text{b)} \quad B_0 e(0) = 0,\\
&\text{c)} \quad (I-Q(\tau))e(\tau) = (I-Q(\tau))\left(u_p(\tau) - u_p^*(\tau)\right)\\
&\qquad\qquad + \left(Q(\tau)-Q^*(\tau)\right)\left(u_p^*(\tau) - \omega_a(\tau)\right)\\
&\qquad\qquad \equiv \Delta(t).
\end{aligned} \tag{3.20}
$$

Note that $\Delta(\tau)$, by construction, is in the range of $I-Q(\tau)$. (We assume, of course, that $\omega_a(x)$ exists.) Therefore we have:

$$(I-Q(\tau))\Delta(\tau) = \Delta(\tau).$$

We now plug into (3.18) to obtain:

$$\|e(x)\| \leqq \left(K_+(x,\tau) + K_-(x,0)K_\phi K_0 K_+(0,\tau)\right)\|\Delta(\tau)\|. \tag{3.21}$$

Further specializing to the case of an exponential dichotomy this becomes:

$$\|e(x)\| \leqq \left(N_+ e^{\alpha_+(x-\tau)} + N_- e^{-\alpha_- x} K_\phi K_0 N_+ e^{-\alpha_+ \tau}\right) \|\Delta(\tau)\|. \tag{3.22}$$

That is, the large part of the error decays exponentially off the artificial boundary.

4. Problems with constant tails. In this section we restrict ourselves to problems which are autonomous in x for x sufficiently large. That is, we assume there exists τ such that:

$$A(x) \equiv A_\infty, \qquad x \geqq \tau. \tag{4.1}$$

We also require that the constant coefficient problem in the tail be separable. That is, we require that a complete spectral representation be associated with A_∞:

Assumption 4.2. There exists a countable set of pairs, (λ_n, u_n), with λ_n a complex number, $u_n \in \mathscr{B}$ and 0 not an accumulation point of $\{\lambda_n\}$ and there exist adjoint pairs, (λ_n^*, v_n), with $v_n \in \text{Dual}(\mathscr{B})$, satisfying

$$\begin{aligned} &\text{i)} \quad A_\infty u_n = \lambda_n u_n; \\ &\text{ii)} \quad A_\infty^* v_n = \lambda_n^* v_n; \\ &\text{iii)} \quad (v_m, u_n) = \delta_{mn}. \end{aligned} \tag{4.2}$$

Furthermore, any function $u \in \mathscr{B}$ can be uniquely written in the form:

$$u = \sum_{n=1}^{\infty} c_n u_n, \qquad c_n = (v_n, u). \tag{4.3}$$

Using the eigenfunction expansions defined above, it is easy to write down conditions for the existence of dichotomies for the constant problem as well as representations of the various operators discussed in the preceding sections. In particular we have the following theorem, whose proof follows immediately from the (formal) solution of the Cauchy problem in terms of the eigenfunction expansions. (For the details of these see Hagstrom [7].)

THEOREM 4.4. a) *If all eigenvalues, λ_n, of A_∞ are bounded away from the imaginary axis, then the homogeneous problem associated with A_∞ has an exponential dichotomy with spaces*

$$\begin{aligned} \mathscr{B}_+ &\equiv \text{span}\{u_i : \text{Re}\lambda_i > 0\}; \\ \mathscr{B}_- &\equiv \text{span}\{u_i : \text{Re}\lambda_i < 0\}. \end{aligned} \tag{4.4}$$

The exponents, $\alpha_\pm$, are given by:

$$\begin{aligned} \alpha_+ &= \underset{\text{Re}\lambda_i > 0}{\text{g.l.b.}} |\text{Re}\lambda_i|; \\ \alpha_- &= \underset{\text{Re}\lambda_i < 0}{\text{g.l.b.}} |\text{Re}\lambda_i|. \end{aligned} \tag{4.5}$$

b) *Let $\mathscr{B}_\pm$ be defined as above and let $\mathscr{B}_0$ be given by*:

$$\mathscr{B}_0 \equiv \text{span}\{u_i : \text{Re}\lambda_i = 0\}. \tag{4.6}$$

Let $\mathscr{B}_0^+ \oplus \mathscr{B}_0^-$ be any direct sum decomposition of $\mathscr{B}_0$. Then an ordinary dichotomy is induced by the spaces $\mathscr{B}_+ \oplus \mathscr{B}_0^+$ and $\mathscr{B}_- \oplus \mathscr{B}_0^-$.

We note that by the conclusions of part (b), there can be many ordinary dichotomies associated with a problem whose operator has eigenvalues with zero real part. Which of these is the right one to use for the boundary condition depends on the boundary operator at infinity, B_∞. Representations of the solution operator, S, are also easy to obtain.

The theorem above can be applied to the example of §1, problem (1.1). Rewriting the problem in first order form according to transformation (2.4), the operator A_∞ is given by:

$$A_\infty = \begin{pmatrix} 0 & -\nabla_y^2 - a_\infty \\ 1 & 0 \end{pmatrix}.$$

Its eigenvalues are given by $\pm\lambda_n$ and $\pm i\alpha_n$, defined by the reduced eigenvalue problem (1.3) through equation (1.4). If (1.3) had no positive eigenvalues, the problem in the tail would have an exponential dichotomy. In the case of an ordinary dichotomy, the boundary condition (1.7a) corresponds to the choice:

$$\mathcal{B}_0^+ \equiv \mathcal{B}_0, \qquad \mathcal{B}_0^- \equiv \varnothing .$$

If, instead of (1.1d), some other condition was imposed (for example a radiation condition) this choice would change. We note that using the integral representation of the boundary condition, (3.10), the condition that the inhomogeneous term vanish in the tail can be replaced by an integrability assumption. The boundary condition, (1.7), is then replaced by:

$$\frac{1}{2}\begin{pmatrix} 1 & \lambda_n \\ \frac{1}{\lambda_n} & 1 \end{pmatrix}\begin{pmatrix} c_n'(\tau) \\ c_n(\tau) \end{pmatrix} = -\frac{1}{2}\int_\tau^\infty ds e^{-\lambda_n(s-\tau)}\begin{pmatrix} 1 & \lambda_n \\ \frac{1}{\lambda_n} & 1 \end{pmatrix}\begin{pmatrix} f_n(s) \\ 0 \end{pmatrix}$$

which implies

$$c_n'(\tau) = -\lambda_n c_n(\tau) - \int_\tau^\infty ds e^{-\lambda_n(s-\tau)} f_n(s), \qquad n = m+1,\ m+2, \cdots . \tag{4.7}$$

For the imaginary eigenvalues we have:

$$\begin{pmatrix} c_n'(\tau) \\ c_n(\tau) \end{pmatrix} = -\frac{1}{2}\int_\tau^\infty ds \left\{ e^{-i\alpha_n(s-\tau)}\begin{pmatrix} 1 & i\alpha_n \\ \frac{1}{i\alpha_n} & 1 \end{pmatrix} + e^{i\alpha_n(s-\tau)}\begin{pmatrix} 1 & -i\alpha_n \\ \frac{1}{i\alpha_n} & 1 \end{pmatrix}\right\}\begin{pmatrix} f_n(s) \\ 0 \end{pmatrix}$$

which implies

$$\begin{aligned} c_n'(\tau) &= -\int_\tau^\infty \cos[\alpha_n(s-\tau)] f_n(s)\, ds; \\ c_n(\tau) &= \frac{1}{\alpha_n}\int_\tau^\infty \sin[\alpha_n(s-\tau)] f_n(s)\, ds. \end{aligned} \tag{4.8}$$

For a general partial differential equation with a constant tail, the eigenvalue problem of its operator, A_∞, can be reduced to an eigenvalue problem for a partial differential operator. In particular, it's eigenvalues, λ, correspond to solutions of:

$$\left[\sum_{j=0}^{n} P_j\left(y, \frac{\partial}{\partial y}\right)\lambda^j\right] Y(y) = 0, \tag{4.9}$$

coupled with the appropriate boundary conditions. This is the eigenvalue problem associated with the Laplace transform in x of the equation in the tail. We note that in practice it is the reduced eigenvalue problem, (4.9), which we suggest be solved to obtain the boundary conditions. The reduction to first order form is made in an effort to simplify the theory. The use of (4.9) to derive boundary conditions was first suggested by Gustafsson and Kreiss [6].

The completeness of the eigenfunctions of A_∞ depends on the completeness of the eigenfunctions of (4.9). This property does not hold in general and is difficult to check. For a class of elliptic and parabolic problems, Agmon and Nirenberg [1, Thm. 5.8] establish the completeness of the eigenfunctions and generalized eigenfunctions of (4.9) whose eigenvalues have negative real part in the class of solutions which are absolutely integrable along with their first $n-1$ x derivatives. In this case, the solution of (4.9) is guaranteed to yield a representation of the admissible space.

5. Perturbation theory and asymptotic boundary conditions. In the preceding section we found useful representations of the projection operator, $Q(\tau)$, of the admissible space and of the particular solution, $u_p(x)$ for equations of the form (2.1) with constant tails. In the present section we relax this assumption and replace it with:

$$\lim_{x\to\infty} A(x)=A_\infty. \tag{5.1}$$

Equivalently we write:

$$A(x)=A_\infty+B(x), \qquad \lim_{x\to\infty}\|B(x)\|=0. \tag{5.2}$$

Assuming A_∞ has a dichotomy, it is possible to make an asymptotic analysis of the perturbed problem defined by $A(x)$. In particular, we obtain representations of the projector, $Q(\tau)$, into the admissible space. Consider the homogeneous problem in the tail:

$$\begin{aligned} &\text{a)} && \frac{dv}{dx}=A_\infty v+B(x)v,\ x\geqq\tau;\\ &\text{b)} && \lim_{x\to\infty} B_\infty v(x)=0;\\ &\text{c)} && \|u(x)\| \text{ bounded as } x\to\infty. \end{aligned} \tag{5.3}$$

Treating $B(x)v$ as an inhomogeneous term, we have, by (3.10), that $v(x)$ must satisfy:

$$(I-Q_\infty(\tau))v(\tau)=-\int_\tau^\infty S(\tau,p;A_\infty)(I-Q_\infty(p))B(p)v(p)\,dp. \tag{5.4}$$

Also, from (3.10), we have a representation of v which must be valid if v exists;

$$\begin{aligned} v(x)=S(x,\tau;A_\infty)Q_\infty(\tau)v(\tau)+\int_\tau^x S(x,p;A_\infty)Q_\infty(p)B(p)v(p)\,dp\\ -\int_x^\infty S(x,p;A_\infty)(I-Q_\infty(p))B(p)v(p)\,dp. \end{aligned} \tag{5.5}$$

Let any $\xi_0\in\mathscr{A}_\infty(\tau)$ be given and replace $Q_\infty(\tau)v(\tau)$ in (5.5) by ξ_0. If the following condition holds:

$$\begin{aligned} \sup_{x\geqq\tau}\Big\|\Big[\int_\tau^x S(x,p;A_\infty)Q_\infty(p)B(p)\,dp\\ -\int_x^\infty S(x,p;A_\infty)(I-Q_\infty(p))B(p)\,dp\Big]\Big\|=K<1; \end{aligned} \tag{5.6}$$

then the contraction mapping theorem can be used to establish the existence of a unique bounded solution to equation (5.5), $v(x;\xi_0)$. Furthermore, we clearly have that:

$$\begin{aligned} &Q_\infty v(\tau;\xi_0)=\xi_0;\\ &(I-Q_\infty(\tau))v(\tau;\xi_0)=-\int_\tau^\infty S(\tau,p;A_\infty)(I-Q_\infty(p))B(p)v(p;\xi_0)\,dp. \end{aligned} \tag{5.7}$$

Hence, whenever (5.6) is valid, we can find, for any $\xi_0\in\mathscr{A}_\infty(\tau)$, a unique element, $v(\tau;\xi_0)$, of $\mathscr{A}(\tau)$. A projector into $\mathscr{A}(\tau)$ is given implicitly by (5.7):

$$Q(\tau)\xi=Q_\infty(\tau)\xi-\int_\tau^\infty dpS(\tau,p;A_\infty)(I-Q_\infty(p))B(p)v(p;Q_\infty(\tau)\xi). \tag{5.8}$$

These conditions lead us to the following theorem:

THEOREM 5.9. *We suppose that either the unperturbed problem has an ordinary dichotomy and* $\|B(x)\|$ *is integrable or that the unperturbed problem has an exponential dichotomy. Then, for* τ *sufficiently large, a unique solution,* $v(x;\xi_0)$, *exists for any* $\xi_0\in\mathscr{A}_\infty(\tau)$ *and* (5.8) *is valid.*

Proof. It is only necessary to satisfy (5.6). In the first case we have:

$$K\leqq(N_+ + N_-)\int_\tau^\infty \|B(x)\|\,dx;$$

while in the second we have:

$$K\leqq\left(\frac{N_+}{\alpha_+}+\frac{N_-}{\alpha_-}\right)\max_{x\geqq\tau}\|B(x)\|.$$

For both cases, the assumptions on B allow us to make the right-hand sides arbitrarily small by choosing τ sufficiently large, completing the proof.

The contraction mapping solution of (5.5) leads to a natural iterative procedure for the approximation of $v(x;\xi_0)$ and, ultimately, of the operator Q. We let:

$$\begin{aligned} &v^{(0)}(x;\xi_0)=S(x,\tau;A_\infty)\xi_0;\\ &v^{(n+1)}(x;\xi_0)=v^{(0)}(x;\xi_0)+\int_\tau^x dpS(x,p;A_\infty)Q_\infty(p)B(p)v^{(n)}(p;\xi_0)\\ &\qquad\qquad -\int_x^\infty dpS(x,p;A_\infty)(I-Q_\infty(p))B(p)v^{(n)}(p;\xi_0). \end{aligned} \tag{5.9}$$

Then, by the contraction estimates:

$$\|v^{(n)}(x;\xi_0)-v(x;\xi_0)\|\leqq\frac{K^{n+1}}{1-K}\|v^{(0)}(x;\xi_0)\|. \tag{5.10}$$

We define our nth approximation to $A(\tau)$, $Q^{(n)}(\tau)$, by:

(5.11)

$$Q^{(n)}(\tau)\xi=Q_\infty(\tau)\xi-\int_\tau^\infty dpS(\tau,p;A_\infty)(I-Q_\infty(p))B(p)v^{(n-1)}(p;Q_\infty(\tau)\xi).$$

The error due to this approximation is estimated by:

$$\|Q(\tau)\xi-Q^{(n)}(\tau)\xi\|\leqq\frac{K^{n+1}}{1-K}\|v^{(0)}(x;Q_\infty(\tau)\xi)\|. \tag{5.12}$$

(Note: in all cases the norm of a $\mathscr{B}$-valued function of x is taken to be the maximum in x of its $\mathscr{B}$ norms.)

We now apply these results to the case when the constant tail problem has an exponential dichotomy and A_∞ has a complete spectrum. We assume that $B(x)$ has an expansion of the form:

$$B(x)=\frac{1}{x}B^{(1)}+\frac{1}{x^2}B^{(2)}+\cdots. \tag{5.13}$$

(The expansions could easily be carried out for more general forms.) Plugging into the formulas above we have:

(5.14)

$$V^{(1)}(x;\xi)=\sum_{\substack{n\\ \mathrm{Re}\,\lambda_n<0}} c_n e^{\lambda_n(x-\tau)}+\sum_{\substack{n\\ \mathrm{Re}\,\lambda_n<0}}\sum_{\substack{m\\ \mathrm{Re}\,\lambda_m<0}}\int_\tau^x dp e^{\lambda_n(x-p)}B_{nm}(p)c_m e^{\lambda_m(p-\tau)}$$

$$-\sum_{\substack{n\\ \mathrm{Re}\,\lambda_n>0}}\sum_{\substack{m\\ \mathrm{Re}\,\lambda_m<0}}\int_x^\infty dp e^{\lambda_n(x-p)}B_{nm}l(p)c_m e^{\lambda_m(p-\tau)};$$

where

$$c_n=(v_n,\xi),\qquad B_{nm}(x)=(v_n,B(x)u_m). \tag{5.15}$$

Using (5.13) and approximating the integrals using integration by parts yields to within an $O(1/\tau^2)$ error:

(5.16)

$$v^{(1)}(x;\xi)=\sum_{\substack{n\\ \mathrm{Re}\,\lambda_n<0}} u_n c_n e^{\lambda_n(x-\tau)}+\sum_{\substack{n\\ \mathrm{Re}\,\lambda_n<n}}\sum_{\substack{m\\ \mathrm{Re}\,\lambda_m<0\\ \lambda_m\neq\lambda_n}} u_n\frac{B^{(1)}_{nm}}{\lambda_m-\lambda_n}c_m\left(\frac{e^{\lambda_m(x-\tau)}}{x}-\frac{e^{\lambda_n(x-\tau)}}{\tau}\right)$$

$$+\sum_{\substack{n\\ \mathrm{Re}\,\lambda_n<0}} u_n B^{(1)}_{nm}c_n e^{\lambda_n(x-\tau)}\log\left(\frac{x}{\tau}\right)+\sum_{\substack{n\\ \mathrm{Re}\,\lambda_n>0}}\sum_{\substack{m\\ \mathrm{Re}\,\lambda_m<0}} u_n\frac{B^{(1)}_{nm}}{\lambda_m-\lambda_n}c_m\frac{e^{\lambda_m(x-\tau)}}{x}.$$

Putting this expression into (5.11) and approximating the integrals in a similar fashion yields:

$$Q(\tau)\xi=\sum_{\substack{n\\ \mathrm{Re}\,\lambda_n<0}} c_n u_n+\sum_{\substack{n\\ \mathrm{Re}\,\lambda_n>0}}\sum_{\substack{m\\ \mathrm{Re}\,\lambda_m<0}} u_n\left(B^{(1)}_{nm}+\frac{1}{\tau}B^{(2)}_{nm}\right)c_m\frac{1}{(\lambda_m-\lambda_n)\tau}$$

$$+\sum_{\substack{n\\ \mathrm{Re}\,\lambda_n>0}}\sum_{\substack{m\\ \mathrm{Re}\,\lambda_m<0}} u_n B^{(1)}_{nm}c_m\frac{1}{(\lambda_m-\lambda_n)^2\tau^2}$$

$$-\sum_{\substack{n\\ \mathrm{Re}\,\lambda_n>0}}\sum_{\substack{j\\ \mathrm{Re}\,\lambda_j<0}}\sum_{\substack{m\\ \mathrm{Re}\,\lambda_m<0}} u_n B^{(1)}_{nj}B^{(1)}_{jm}c_m\frac{1}{\tau^2(\lambda_m-\lambda_n)(\lambda_j-\lambda_n)} \tag{5.17}$$

$$+\sum_{\substack{n\\ \mathrm{Re}\,\lambda_n>0}}\sum_{\substack{j\\ \mathrm{Re}\,\lambda_j>0}}\sum_{\substack{m\\ \mathrm{Re}\,\lambda_m<0}} u_n B^{(1)}_{nj}B^{(1)}_{jm}c_m\frac{1}{\tau^2(\lambda_m-\lambda_j)(\lambda_m-\lambda_n)}+O\left(\frac{1}{\tau^3}\right);$$

$$c_n=(v_n,\xi).$$

The generality of the expansion given above makes its automatic computation a real possibility. Note that the expansion is equivalent to the one obtained by Jepson and Keller [10] for ordinary differential equations.

Formula (5.17) can be applied to the Laplacian example, (1.1), where the potential $a(\mathbf{y})$ is replaced by:

$$a(x,\mathbf{y}) = a_0(\mathbf{y}) + \frac{1}{x}a_1(\mathbf{y}) + \frac{1}{x^2}a_2(\mathbf{y}) + \cdots. \tag{5.18}$$

Then, the matrix elements $B_{nm}^{(i)}$ are given by:

$$B_{nm}^{(i)} = \frac{1}{2\lambda_n}\int_{\Omega} d\mathbf{y}\, Y_n(\mathbf{y}) Y_m(\mathbf{y}) a_i(\mathbf{y}). \tag{5.19}$$

Expansions of a particular solution can be derived in a similar manner. Let $u_\infty(x)$ be any particular solution of the unperturbed problem. Then, a solution of the integral equation:

$$\begin{aligned} u(x) = u_\infty(x) &+ \int_\tau^x 4k\, S(x,p) Q_\infty(p) B(p) u(p)\, dp \\ &- \int_x^\infty S(x,p)(I - Q_\infty(p)) B(p) u(p)\, dp \end{aligned} \tag{5.20}$$

is a particular solution of the perturbed problem. Given the inequality (5.6), a unique bounded solution of (5.20) exists by the contraction mapping theorem. It can be approximated by an iterative process analogous to the one described by (5.9). Perturbations of the inhomogeneous term could also be included.

Finally, we note that (5.17) is valid for some problems which do not satisfy (5.6). An important example is afforded by the exterior Helmholtz problem in two dimensions. The equation in the tail is:

$$\frac{\partial^2 u}{\partial r^2} + \frac{1}{r}\frac{\partial u}{\partial r} + \frac{1}{r^2}\frac{\partial^2 u}{\partial \theta^2} + k^2 u = 0, \qquad r \in [\tau, \infty), \quad \theta \in [0, 2\pi) \tag{5.21}$$

together with boundary conditions

$$\begin{aligned} &\text{a)} \quad u \text{ periodic in } \theta; \\ &\text{b)} \quad \lim_{r\to\infty} r^{1/2}\left(\frac{\partial u}{\partial r} - iku\right) = 0. \end{aligned} \tag{5.22}$$

Rewritten in first order form these become:

$$\begin{aligned} &\text{a)} \quad \frac{\partial}{\partial r}\begin{pmatrix}\omega\\ u\end{pmatrix} + \begin{pmatrix} 0 & k^2 \\ -1 & 0\end{pmatrix}\begin{pmatrix}\omega\\ u\end{pmatrix} + \frac{1}{r}\begin{pmatrix}1 & 0\\ 0 & 0\end{pmatrix}\begin{pmatrix}\omega\\ u\end{pmatrix} \\ &\qquad\qquad + \frac{1}{r^2}\begin{pmatrix} 0 & \dfrac{\partial^2}{\partial\theta^2} \\ 0 & 0\end{pmatrix}\begin{pmatrix}\omega\\ u\end{pmatrix} = \begin{pmatrix}0\\ 0\end{pmatrix}; \\ &\text{b)} \quad \begin{pmatrix}\omega\\ u\end{pmatrix} \text{ periodic in } \theta; \\ &\text{c)} \quad \lim_{r\to\infty} r^{1/2}(\omega - iku) = 0. \end{aligned} \tag{5.23}$$

There are two obstacles to the application of the preceding theory to problem (5.23). The first is that the perturbation

$$\frac{1}{r^2}\begin{pmatrix} 0 & \dfrac{\partial^2}{\partial\theta^2} \\ 0 & 0 \end{pmatrix}$$

is apparently unbounded. The second is that the perturbation

$$\frac{1}{r}\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$$

is nonintegrable while the limiting equation,

$$\frac{\partial}{\partial r}\begin{pmatrix} \omega \\ u \end{pmatrix}+\begin{pmatrix} 0 & k^2 \\ -1 & 0 \end{pmatrix}\begin{pmatrix} \omega \\ u \end{pmatrix}=\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \tag{5.24}$$

has an ordinary rather than an exponential dicohotomy. Nonetheless, it is possible to apply formula (5.17), or any higher order approximation to the boundary condition, to this problem. (It is necessary, of course, to identify the imaginary eigenvalue ik with eigenvalues with negative real part and $-ik$ with eigenvalues with positive real part when applying the formulas.) The resulting boundary condition is:

(5.25)

$$\begin{aligned}(I-Q(\tau))\begin{pmatrix} \omega(\tau) \\ u(\tau) \end{pmatrix} &= \frac{1}{2}\begin{pmatrix} 1 & -ik \\ -\dfrac{1}{ik} & 1 \end{pmatrix}\begin{pmatrix} \omega(\tau) \\ u(\tau) \end{pmatrix} \\ &\quad -\frac{1}{2ik\tau}\left(\frac{1}{2}+\frac{1}{2ik\tau}\frac{\partial^2}{\partial\theta^2}\right)\begin{pmatrix} -\dfrac{1}{2} & \dfrac{-ik}{2} \\ \dfrac{1}{2ik} & \dfrac{1}{2} \end{pmatrix}\begin{pmatrix} \omega(\tau) \\ u(\tau) \end{pmatrix} \\ &\quad +\frac{1}{4k^2\tau^2}\cdot\frac{1}{2}\begin{pmatrix} -\dfrac{1}{2} & \dfrac{-ik}{2} \\ \dfrac{1}{2ik} & \dfrac{1}{2} \end{pmatrix}\begin{pmatrix} \omega(\tau) \\ u(\tau) \end{pmatrix} \\ &\quad +\frac{1}{4k^2\tau^2}\cdot\frac{1}{4}\begin{pmatrix} -\dfrac{1}{2} & \dfrac{-ik}{2} \\ \dfrac{1}{2ik} & \dfrac{1}{2} \end{pmatrix}\begin{pmatrix} \omega(\tau) \\ u(\tau) \end{pmatrix} \\ &\quad -\frac{1}{4k^2\tau^2}\cdot\frac{1}{4}\begin{pmatrix} -\dfrac{1}{2} & \dfrac{-ik}{2} \\ \dfrac{1}{2ik} & \dfrac{1}{2} \end{pmatrix}\begin{pmatrix} \omega(\tau) \\ u(\tau) \end{pmatrix}+O\left(\frac{1}{\tau^3}\right)=\begin{pmatrix} 0 \\ 0 \end{pmatrix},\end{aligned}$$

which can be written:

$$\frac{\partial u}{\partial r}(\tau,\theta)=iku(\tau,\theta)-\frac{1}{2\tau}u(\tau,\theta)-\frac{1}{2ik\tau^2}\frac{\partial^2 u}{\partial\theta^2}(\tau,\theta)-\frac{1}{8ik\tau^2}u(\tau,\theta). \tag{5.26}$$

The validity of (5.26) can be established by other means. See, for example, Bayliss, Gunzburger and Turkel [2]. We note that the error depends on higher θ derivatives of u.

6. Nonlinear problems. In this final section we apply the perturbation theory to nonlinear problems. We restrict ourselves to abstract problems of the form:

$$\begin{aligned} &\text{a)} \quad \frac{du}{dx} = F(u), \qquad x \geqq \tau; \\ (6.1)\quad &\text{b)} \quad \lim_{x\to\infty} u(x) = u_\infty; \\ &\text{c)} \quad F(u_\infty) = 0; \end{aligned}$$

where $u(x)$ is an element of some Banach space, $\mathscr{B}$, and F is a nonlinear operator with domain and range in $\mathscr{B}$. Letting $v = u - u_\infty$, we rewrite (6.1):

$$\begin{aligned} &\text{a)} \quad \frac{dv}{dx} = F_u(u_\infty)v + R(v), \qquad x \geqq \tau; \\ (6.2)\quad &\text{b)} \quad \lim_{x\to\infty} v(x) = 0; \\ &\text{c)} \quad R(v) \equiv F(u_\infty + v) - F_u(u_\infty)v. \end{aligned}$$

One approach to the solution of (6.1) or (6.2) would be Newton's method. Then, the theory of the preceding sections could be applied at each stage of the iteration. We, however, choose to work directly with (6.2), deriving exact boundary conditions which can be approximated by the methods of §5.

We generalize the notion of an admissible set (Definition 2.5) to be applicable to (6.2). Note that it is no longer an affine subset of $\mathscr{B}$. Central to our analysis is the behavior of solutions to the linearized problem in the tail:

$$\frac{d\omega}{dx} = F_u(u_\infty)\omega, \qquad x \geqq \tau. \tag{6.3}$$

Treating the nonlinearity, $R(v)$, as an inhomogeneous term leads to the following equations for v, which are analogous to (5.4) and (5.5);

$$(I - Q_\infty(\tau))v(\tau) = -\int_\tau^\infty S(\tau, p; F_u(u_\infty))(I - Q_\infty(p))R(v(p))\,dp; \tag{6.4}$$

$$\begin{aligned} v(x) = {} & S(x,\tau; F_u(u_\infty))Q_\infty(\tau)v(\tau) + \int_\tau^x S(x,p; F_u(u_\infty))Q_\infty(p)R(v(p))\,dp \\ & - \int_x^\infty S(x,p; F_u(u_\infty))(I - Q_\infty(p))R(v(p))\,dp. \end{aligned} \tag{6.5}$$

Here, Q_∞ projects into the admissible space of the linearized problem (6.3). As in the linear case, the condition that (6.4) and (6.5) be simultaneously solvable is viewed as a condition for the admissibility of $v(\tau)$.

Following the derivation for the linear problem, we let $\xi_0 \in \mathscr{A}_\infty(\tau)$ be given and use a contraction argument to establish the existence of a solution to the integral equation, (6.5), with $Q_\infty(\tau)v(\tau)$ replaced by ξ_0. Due to the nonlinearity, some additional assumptions are needed:

Assumption 6.6. a) There exists $\delta > 0$ such that if u_1, $u_2 \in \mathscr{B}$ and $\|u_i\| \leqq \delta$, $i = 1, 2$, then

$$\begin{aligned} \sup_{x\geqq\tau} \Big\| & \int_\tau^x S(x,p; F_u(u_\infty))Q_\infty(p)(R(u_1) - R(u_2))\,dp \\ & - \int_x^\infty S(x,p; F_u(u_\infty))(I - Q_\infty(p))(R(u_1) - R(u_2))\,dp \Big\| \\ & \leqq K\|u_1 - u_2\|, \qquad K < 1. \end{aligned}$$

b) There exists $\delta_1 > 0$ such that if $u \in \mathscr{B}$ and $\|u\| < \delta$, then

$$\sup_{x \geqq \tau} \left\| \int_\tau^x S(x,p;F_u(u_\infty))Q_\infty(p)R(u)\,du - \int_x^\infty S(x,p;F_u(u_\infty))(I-Q_\infty(p))R(u)\,dp \right\| \leqq \delta - \delta_1.$$

c)

$$\sup_{x \geqq \tau} \|S(x,\tau;F_u(u_\infty))\xi_0\| < \delta_1.$$

Given these, a solution to (6.5) is guaranteed by the contraction mapping theorem. Denoting this solution by $v(x;\xi_0)$, an exact boundary condition, valid for small boundary data, can be written down from (6.4):

(6.6)

$$(I-Q_\infty(\tau))v(\tau) = -\int_\tau^\infty S(\tau,p;F_u(u_\infty))(I-Q_\infty(p))R(v(p;Q_\infty(\tau)v(\tau)))\,dp.$$

An approximation to (6.6) can be obtained from an iterative approximation to the solution of (6.5):

(6.7)

a) $v^{(0)}(x;\xi_0) = S(x,\tau;F_u(u_\infty))\xi_0,$

b) $v^{(n+1)}(x;\xi_0) = v^{(0)}(x;\xi_0) + \int_\tau^x S(x,p;F_u(u_\infty))Q_\infty(p)R(v^{(n)}(p;\xi_0))\,dp - \int_x^\infty S(x,p;F_u(u_\infty))(I-Q_\infty(p))R(v^{(n)}(p;\xi_0))\,dp.$

The nth approximation to the boundary condition is, then, given by:

(6.8)

$$(I-Q_\infty(\tau))v(\tau) = -\int_\tau^\infty dpS(\tau,p;F_u(u_\infty))(I-Q_\infty(p))R(v^{(n)}(p;Q_\infty(\tau)v(\tau))).$$

Error estimates follow as in the linear case and will be proportional to $K^{n+1}\|v^{(0)}\|$ which, in turn, we expect to be proportional to $\|v(\tau)\|^{n+2}$. Note that R will often be given as an expansion:

$$R(v) \sim \frac{1}{2}F_{uu}(u_\infty)vv + \frac{1}{6}F_{uuu}(u_\infty)vvv + \cdots. \tag{6.9}$$

We take as many terms in this expansion when evaluating the integrals as is consistent with the number of terms in (6.7) we intend to retain.

Assume now that the linearized operator, $F_u(u_\infty)$, has a complete spectrum. Then, in order to satisfy part (a) of assumption (6.6), it is necessary to assume that there is an exponential dichotomy. From (6.9) we derive the following representation of $R(v)$ in terms of the eigenfunctions of $F_u(u_\infty)$:

(6.10)

a) $v = \sum_{n=1}^\infty c_n u_n,\ c_n = (v_n, u);$

b) $R(v) = \sum_{n=1}^\infty \gamma_n(v)u_n,\ \gamma_n(v) \sim \sum \alpha_{ij}^{(n)} c_i c_j + \sum_{i,j,k} \beta_{ijk}^{(n)} c_i c_j c_k + \cdots;$

c) $\alpha_{ij}^{(n)} = \left(v_n, \frac{1}{2}F_{uu}u_i u_j\right),\ \beta_{ijk}^{(n)} = \left(v_n, \frac{1}{6}F_{uuu}u_i u_j u_k\right), \cdots.$

The function $v^{(1)}(x;\xi)$ is given by:

(6.11)

$$
\begin{aligned}
v^{(1)}(x;\xi) = & \sum_{\substack{n \\ \mathrm{Re}\lambda_n<0}} u_n c_n e^{\lambda_n(x-\tau)} \\
& + \sum_{\substack{n \\ \mathrm{Re}\lambda_n<0}} \sum_{\substack{i \\ \mathrm{Re}\lambda_i<0 \\ \lambda_i+\lambda_j\neq\lambda_n}} \sum_{\substack{j \\ \mathrm{Re}\lambda_j<0}} u_n \alpha_{ij}^n \frac{c_i c_j}{(\lambda_i+\lambda_j-\lambda_n)} \left(e^{(\lambda_i+\lambda_j)(x-\tau)} - e^{\lambda_n(x-\tau)}\right) \\
& + \sum_{\substack{n \\ \mathrm{Re}\lambda_n<0}} \sum_{\substack{i \\ \mathrm{Re}\lambda_i<0 \\ \lambda_i+\lambda_j=\lambda_n}} \sum_{\substack{j \\ \mathrm{Re}\lambda_j<0}} u_n \alpha_{ij}^n c_i c_j e^{\lambda_n(x-\tau)}(x-\tau) \\
& + \sum_{\substack{n \\ \mathrm{Re}\lambda_n>0}} \sum_{\substack{i \\ \mathrm{Re}\lambda_i<0}} \sum_{\substack{j \\ \mathrm{Re}\lambda_j<0}} u_n \alpha_{ij}^n c_i c_j \frac{1}{\lambda_i+\lambda_j-\lambda_n} e^{(\lambda_i+\lambda_j)(x-\tau)} + O\left(\|\xi\|^3\right); \\
c_n = & (v_n, \xi).
\end{aligned}
$$

This yields the following approximation to the boundary condition, which we write in terms of the expansion coefficients. Here, n is such that $\mathrm{Re}\lambda_n > 0$.

(6.12)

$$
\begin{aligned}
c_n = & \sum_{\substack{i \\ \mathrm{Re}\lambda_i<0}} \sum_{\substack{j \\ \mathrm{Re}\lambda_j<0}} \alpha_{ij}^n \frac{c_i c_j}{\lambda_i+\lambda_j-\lambda_n} + \sum_{\substack{i \\ \mathrm{Re}\lambda_i<0}} \sum_{\substack{j \\ \mathrm{Re}\lambda_j<0}} \sum_{\substack{k \\ \mathrm{Re}\lambda_k<0}} \beta_{ijk}^n c_i c_j c_k \frac{1}{\lambda_i+\lambda_j+\lambda_k-\lambda_n} \\
& - \sum_{\substack{i \\ \mathrm{Re}\lambda_i<0}} \sum_{\substack{j \\ \mathrm{Re}\lambda_j<0}} \sum_{\substack{k \\ \mathrm{Re}\lambda_k<0}} \sum_{\substack{l \\ \mathrm{Re}\lambda_l<0}} \left\{\alpha_{ij}^n \alpha_{kl}^j + \alpha_{ji}^n \alpha_{kl}^j\right\} c_i c_k c_l \frac{1}{(\lambda_k+\lambda_l+\lambda_i-\lambda_n)(\lambda_i+\lambda_j-\lambda_n)} \\
& + \sum_{\substack{i \\ \mathrm{Re}\lambda_i<0}} \sum_{\substack{j \\ \mathrm{Re}\lambda_j>0}} \sum_{\substack{k \\ \mathrm{Re}\lambda_k<0}} \sum_{\substack{l \\ \mathrm{Re}\lambda_l<0}} \left\{\alpha_{ij}^n \alpha_{kl}^j + \alpha_{ji}^n \alpha_{kl}^j\right\} c_i c_k c_l \frac{1}{(\lambda_k+\lambda_l+\lambda_i-\lambda_n)(\lambda_k+\lambda_l-\lambda_j)}.
\end{aligned}
$$

This general formula can be applied, for example, to nonlinear elliptic problems of the form:

$$
\begin{aligned}
&\text{a)} \quad \nabla^2 u = f(u, \mathbf{y}), \qquad (x, \mathbf{y}) \in [\tau, \infty) \times \Omega; \\
&\text{b)} \quad B_\Omega u = 0, \qquad \mathbf{y} \in \partial\Omega; \\
&\text{c)} \quad \lim_{x\to\infty} u(x, \mathbf{y}) = u_\infty(\mathbf{y});
\end{aligned} \tag{6.13}
$$

where $u_\infty(\mathbf{y})$ satisfies;

$$
\begin{aligned}
&\text{a)} \quad \nabla_{\mathbf{y}}^2 u_\infty = f(u_\infty, \mathbf{y}), \qquad \mathbf{y} \in \Omega; \\
&\text{b)} \quad B_\Omega u_\infty = 0, \qquad \mathbf{y} \in \partial\Omega.
\end{aligned} \tag{6.14}
$$

The linearized equation in the tail is given by:

$$
\nabla^2 v - f_u(u_\infty, \mathbf{y}) v = 0; \tag{6.15}
$$

which is of the form analyzed in §3. The condition that (6.15) have an exponential dichotomy is that all eigenvalues, α_n, of the problem

$$(6.16)\quad \begin{aligned} &\text{a)} \quad \nabla_{\mathbf{y}}^2 Y_n - f_u(u_\infty, \mathbf{y}) Y_n = \alpha_n Y_n, \qquad \mathbf{y} \in \Omega; \\ &\text{b)} \quad B_\Omega Y_n = 0, \qquad \mathbf{y} \in \partial\Omega; \end{aligned}$$

be negative. Then, the following boundary condition can be derived from (6.12):

$$(6.17)\quad c_n' = -\lambda_n c_n - \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} \alpha_{ij}^n \frac{c_i c_j}{\lambda_i+\lambda_j+\lambda_n} - \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\sum_{k=1}^{\infty} \beta_{ijk}^n \frac{c_i c_j c_k}{\lambda_i+\lambda_j+\lambda_k+\lambda_n}$$
$$+ \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\sum_{k=1}^{\infty}\sum_{l=1}^{\infty} \alpha_{ij}^n \alpha_{kl}^j \frac{c_i c_k c_l}{\lambda_k+\lambda_l+\lambda_i+\lambda_n}\left(\frac{1}{\lambda_k+\lambda_l+\lambda_j} - \frac{1}{\lambda_i+\lambda_n+\lambda_j}\right);$$
$$n = 1, 2, 3, \cdots.$$

Here we have:

$$(6.18)\quad \begin{aligned} c_i &= \int_\Omega d\mathbf{y}\, Y_i(\mathbf{y}) V(\tau, \mathbf{y}); \\ \lambda_i &= \sqrt{-\alpha_i}\,; \\ \alpha_{ij}^n &= \int_\Omega d\mathbf{y}\, \frac{1}{2} f_{uu}(u_\infty, \mathbf{y}) Y_i(\mathbf{y}) Y_j(\mathbf{y}) Y_n(\mathbf{y}); \\ \beta_{ijk}^n &= \int_\Omega d\mathbf{y}\, \frac{1}{6} f_{uuu}(u_\infty, \mathbf{y}) Y_i(\mathbf{y}) Y_j(\mathbf{y}) Y_k(\mathbf{y}) Y_n(\mathbf{y}). \end{aligned}$$

The quadratic approximation to this condition is used in a numerical computation by the authors in [8].

REFERENCES

[1] S. AGMON AND L. NIRENBERG, *Properties of solutions of ordinary differential equations in Banach space*, Comm. Pure Appl. Math., 16 (1963), pp. 121–239.

[2] A. BAYLISS, M. GUNZBURGER AND E. TURKEL, *Boundary conditions for the numerical solution of elliptic equations in exterior regions*, SIAM J. Appl. Math., 42 (1982), pp. 430–451.

[3] J. BEREZANSKII, *Expansions in eigenfunctions of self-adjoint operators*, Transl. Math. Mon. 17, American Mathematical Society, Providence, RI, 1968.

[4] J. DALETSKIY AND M. KREIN, *Stability of solutions of differential equations in Banach space*, Transl. Math. Mon. 43, American Mathematical Society, Providence, RI, 1974.

[5] F. de HOOG AND R. WEISS, *An approximation method for boundary value problems on infinite intervals*, Computing, 24 (1980), pp. 227–239.

[6] B. GUSTAFSSON AND H.-O. KREISS, *Boundary conditions for time dependent problems with an artificial boundary*, J. Comp. Phys., 30 (1979), pp. 333–351.

[7] T. HAGSTROM, *Reduction of unbounded domains to bounded domains for partial differential equation problems*, Ph. D. thesis, California Inst. of Technology, Pasadena, 1983.

[8] T. HAGSTROM AND H. B. KELLER, *The numerical solution of semi-linear elliptic problems in unbounded cylindrical domains*, to appear.

[9] _____, *The numerical calculation of traveling wave solutions of parabolic equations*, to appear.

[10] A. JEPSON AND H. B. KELLER, *Asymptotic boundary conditions for ordinary differential equations*, to appear.

[11] H. B. KELLER AND M. LENTINI, *Boundary value problems over semi-infinite intervals and their numerical solution*, SIAM J. Numer. Anal., 17 (1980), pp. 577–604.

[12] P. MARKOWICH *A theory for the approximation of solutions of boundary value problems on infinite intervals*, this Journal, 13 (1982), pp. 484–513.

SIAM J. MATH. ANAL.
Vol. 17, No. 2, March 1986

005

CHEMICAL SURFACE REACTIONS AND NONLINEAR STABILITY BY THE METHOD OF ENERGY*

CAROL L. McTAGGART† AND BRIAN STRAUGHAN‡

Abstract. By means of energy stability theory, the nonlinear stability of a two-component reactive fluid, composed of the dimer A_2 and the monomer A, confined between two infinite parallel plates and subject to the surface catalyzed reaction ($A_2 \rightleftharpoons 2A$), is analysed. The stability boundary is calculated in the cases (a) when the catalytic plate is conducting, in which case the linear and nonlinear energy parameter boundaries coincide and (b) when the plate is insulating, for which a global stability criterion is found.

1. Introduction. The convective instability which results when a chemically inert fluid in a gravitational field is heated from below has long been recognized as a problem of crucial importance in many fields of fluid mechanics. More recently, however, the effects caused by finer details such as chemical reactions or phase changes have been shown to play an important role on convection in specific applications, see e.g., Bdzil and Frisch [2], [3], Loper and Roberts [8]. The purpose of this work is to provide a *nonlinear* stability analysis for the conduction-diffusion Bénard problem in which the upper surface is stress free while the lower surface experiences a catalyzed chemical reaction.

Throughout we confine attention to the effects which the heterogeneous surface catalyzed reaction $A_2 \rightleftharpoons 2A$ may have on the hydrodynamic stability of the fluid mixture containing the dimer A_2 and monomer A, where we suppose the fluid is only slightly removed from chemical equilibrium and contained in the layer between the surfaces $z=0$ and $z=d$, the lower surface being catalytic. Bdzil and Frisch [2], [3] draw attention to the application of such problems to the dissociation of oxygen, hydrogen or nitrogen near a hot surface such as occurs in the vicinity of the gas-solid interface of a space vehicle upon re-entry into the earth's atmosphere. We would, however, anticipate future applications of our results in laboratory controlled experiments.

The model we adopt is that of Bdzil and Frisch [2], [3] who restrict attention to linearized instability analyses. The situation is described by a Newtonian fluid model, to which a Boussinesq approximation has been applied, in which the basic fields are those of velocity $\mathbf{v}$, pressure p, temperature T and degree of dissociation α ($=$ fraction of pure monomers present). The novelty of the problem is best described in terms of the heat flux $\mathbf{q}$ and mass flux $\mathbf{J}$ (local flux of α) in terms of which the conditions to be satisfied at the catalyzed boundary ($z=0$) are

$$\mathbf{J}\cdot\mathbf{k}=R, \quad \text{with either (A) } T=T_0, \quad \text{or (B) } \mathbf{q}\cdot\mathbf{k}=0, \tag{1.1}$$

where T_0 is a prescribed temperature, $\mathbf{k}$ is the vector $(0,0,1)$ and R is the rate at which the monomer is formed by the surface reaction. The reaction rate is, in fact, taken to be

*Received by the editors October 17, 1983.

†Department of Mathematics, University of Glasgow, Glasgow G12 8QW, Scotland. Present address, Department of Theoretical Mechanics, University of Nottingham, University Park, Nottingham NG7 2RD, England. The work of this author was carried out while she held a Research Studentship of the Science and Engineering Research Council of Great Britain.

‡Department of Mathematics, University of Glasgow, Glasgow G12 8QW, Scotland.

linear in α, i.e.,

$$R = -\sigma(\alpha - \alpha^e), \tag{1.2}$$

where α^e is the equilibrium value of α on the surface.

The solutions to the model in the steady state ($\mathbf{v}^s \equiv \mathbf{0}$) are given by [2, (9)] and for α, T are (the pressure p^s is also easily calculated but not given here as it is not required)

$$\alpha^s(z) = \alpha_d + \gamma(d-z), \qquad T^s(z) = T_d + \beta(d-z), \tag{1.3}$$

where α_d, T_d are the prescribed values of α and T on $z=d$, and γ, β are constants which for positive constants K_1, K_2 (given in [2]) may be written as

$$\gamma = -\sigma K_1(\alpha_d - \alpha^e), \tag{1.4}$$

$$\beta = \begin{cases} (T_0 - T_d)/d & \text{if (1.1) (A) is adopted,} \\ \sigma K_2(\alpha_d - \alpha^e) & \text{when (1.1) (B) holds.} \end{cases} \tag{1.5}$$

Our object is to investigate the nonlinear stability of $(\mathbf{v}^s, T^s, \alpha^s, p^s)$ and to this end we must determine the governing equations for an arbitrary perturbation $(\mathbf{u}, \theta, \phi, p)$ to this solution. This calculation is routine and so omitted; however, after nondimensionalization the required equations are seen to be (from the governing equations in [2])

$$\begin{aligned} &u_{i,t} + u_j u_{i,j} = -p_{,i} + \Delta u_i + (R\theta + S\phi)k_i, \qquad \begin{cases} \mathrm{Pr}(\theta_{,t} + u_i\theta_{,i}) = \Delta\theta + H_1 R w, \\ \mathrm{Sc}(\phi_{,t} + u_i\phi_{,i}) = \Delta\phi + H_2 S w, \end{cases} \\ &u_{i,i} = 0, \end{aligned} \tag{1.6}$$

which are to be solved in the layer $\mathbb{R}^2 \times (0,1)$.

In (1.6) $\mathbf{u} = (u,v,w)$, Pr, Sc are the Prandtl and Schmidt numbers, $H_1 = \operatorname{sgn}\beta$, $H_2 = \operatorname{sgn}\gamma$ and R^2, S^2 are the Rayleigh and dissociation Rayleigh numbers, respectively.

Finally, the complete boundary conditions which the solution to (1.6) must satisfy are:

$$\begin{aligned} &\mathbf{u} = \mathbf{0}, \quad z=0, \qquad \frac{\partial u}{\partial z} = \frac{\partial v}{\partial z} = w = 0, \quad z=1, \\ &\theta = \phi = 0, \quad z=1, \\ &\frac{\partial \phi}{\partial z} = r\phi, \quad z=0, \qquad \text{either (A) } \theta = 0, \ z=0, \\ &\qquad\qquad\qquad\qquad\quad \text{or} \quad \text{(B) } \frac{\partial \theta}{\partial z} = -s\phi, \ z=0, \end{aligned} \tag{1.7}$$

where r, s are positive nondimensional surface reaction numbers. In accordance with the observed cellular structure of Bénard instabilities in the presence of a free surface, we shall restrict attention to perturbations $(\mathbf{u}, \theta, \phi, p)$ which are periodic in x and y. A typical period cell in the nondimensional layer will be denoted by Ω and that part of the surface $z=0$ which forms part of the boundary of Ω will be denoted by Γ.

It is convenient to refer to the two situations (A) and (B) above, which correspond to prescribed temperature or thermal insulation, as cases (A) and (B).

2. Nonlinear energy stability for case (A). Mathematically, case (A) and case (B) are very different. In this section we examine case (A) when the applied temperature gradient is destabilizing, i.e., $H_1 = 1$, and when $\alpha_d < \alpha^e$, i.e., $H_2 = 1$. Under these conditions system (1.6) together with the appropriate boundary conditions (1.7) possesses the

desirable property that the linear operator associated with the linear time independent equations is symmetric with respect to the usual inner product on the Hilbert space $(L^2(\Omega))^5$. To verify this $\mathscr{L}$ denote the linear operator defined by the right-hand side of (1.6) and let $D(\mathscr{L})$ be the subset of $(L^2(\Omega))^5$ consisting of C^2 functions whose first three components are solenoidal and which satisfy the boundary conditions of case (A). Then if $(\mathbf{u}^\alpha, \theta^\alpha, \phi^\alpha) = a^\alpha$, $\alpha = 1,2$, are any two elements in $D(\mathscr{L})$, and we denote by $\langle \cdot, \cdot \rangle$ and $D(\cdot, \cdot)$ the inner product on $(L^2(\Omega))^5$ and the Dirichlet integral, a routine calculation shows that

$$\begin{aligned}\langle \mathscr{L}a^1, a^2 \rangle &= -D(\mathbf{u}^1, \mathbf{u}^2) - D(\theta^1, \theta^2) - D(\phi^1, \phi^2) \\ &\quad - r\int_\Gamma \phi^1\phi^2\, dA + R\int_\Omega \left[\theta^1 w^2 + w^1\theta^2\right] dx + S\int_\Omega \left[\phi^1 w^2 + w^1\phi^2\right] dx \\ &= \langle a^1, \mathscr{L}a^2 \rangle,\end{aligned}$$

and the truth of our claim is established.

To determine the nonlinear stability limit we introduce an energy

$$E(t) = \frac{1}{2}\|u\|^2 + \frac{1}{2}\mathrm{Pr}\|\theta\|^2 + \frac{1}{2}\mathrm{Sc}\|\phi\|^2, \tag{2.1}$$

$\|\cdot\|$ denoting the $L^2(\Omega)$ norm, and from (1.6), (1.7) case (A) we may then derive

$$\dot{E} = -\mathscr{D} + I \leqq -\mathscr{D}\left(1 - \max\frac{I}{\mathscr{D}}\right), \tag{2.2}$$

where the maximum is taken over the set of admissible solutions and where

$$\mathscr{D} = D(\mathbf{u}) + D(\theta) + D(\phi), \tag{2.3}$$

$$I = 2R\langle \theta, w \rangle + 2S\langle \phi, w \rangle - r\int_\Gamma \phi^2\, dA. \tag{2.4}$$

Here and for the remainder of the paper $D(\cdot)$ denotes the Dirichlet integral and $\langle \cdot, \cdot \rangle$ is the inner product on $L^2(\Omega)$. Since $\theta = \phi = 0$ on $z = 1$, Poincaré's inequality ensures the existence of a constant $\xi^2 > 0$ such the $\mathscr{D} \geqq \xi^2 E$, and so if

$$\max\frac{I}{D} < 1, \tag{2.5}$$

then (2.2) yields

$$\dot{E} \leqq -\xi^2 E\left(1 - \max\frac{I}{\mathscr{D}}\right),$$

from which it follows that $E \to 0$, $t \to \infty$, and the steady solution of (1.3) is nonlinearly stable. In [5] it is shown that for a symmetric system (2.5) is equivalent to finding the critical values of R and S of the time independent linearized system of equations from which (2.5) arose. In the present context this means we must determine the lowest eigenvalues R, S of the system (1.6), with the nonlinear terms and all time derivative terms set equal to zero, which meet the boundary conditions (1.7) case (A).

This is a convenient point to observe that if the *linearized* system of equations corresponding to (1.6) is studied then the symmetry of $\mathscr{L}$ ensures that the eigenvalues λ_n of linear theory arising from a time dependence like $e^{-\lambda t}$ are all real and so the results for linear stability coincide with those for nonlinear stability. The connection between

the linear and nonlinear results follows from the work of Galdi and Straughan [5]. The importance of this fact is that subcritical instabilities cannot occur and so instability may occur only by stationary convection.

The $\max(I/\mathscr{D})$ problem of determining the criteria for *nonlinear stability* of solution (1.3) then reduces to solving the linear eigenvalue problem

$$\begin{aligned} &\Delta u_i+(R\theta+S\phi)k_i=p_{,i}, \quad \Delta\phi+Sw=0, \\ &\Delta\theta+Rw=0, \end{aligned} \tag{2.6}$$

in Ω, with $\mathbf{u}$ solenoidal, and

$$\begin{aligned} &\mathbf{u}=0, \quad z=0, \qquad \frac{\partial u}{\partial z}=\frac{\partial v}{\partial z}=w=0, \quad z=1, \\ &\theta=\phi=0, \quad z=1, \quad \frac{\partial\phi}{\partial z}=r\phi, \quad \theta=0, \quad z=0. \end{aligned} \tag{2.7}$$

Even this system does not seem solvable analytically because of complications caused by the $z=0$ boundary conditions, although we could solve it numerically. However, the case of constant temperature is of interest and was solved in the linear case by Bdzil and Frisch [2] *when both boundaries are either stress free or fixed.* We are able to present an analytical treatment in the isothermal case and this is given next.

For the isothermal problem (2.6), (2.7) hold with $R\equiv\theta=0$. Due to linearity and the assumed periodicity of the solution a representation in the form $\mathbf{u}(\mathbf{x})=e^{i(kx+my)}\mathbf{u}(z)$ is possible, with a similar form for ϕ,p. Let $u_3(z)=W(z)$, $D=d/dz$ and $a^2=k^2+m^2$, then (2.6), (2.7) reduce to

$$(D^2-a^2)^3W=-a^2S^2W, \qquad z\in(0,1), \tag{2.8}$$

$$\begin{aligned} &W=DW=(D^2-a^2)^2(D-r)W=0, \quad z=0, \\ &W=D^2W=(D^2-a^2)^2W=0, \qquad z=1. \end{aligned} \tag{2.9}$$

Bdzil and Frisch [2] determine the first eigenvalue S_c^2 of (2.8), but with (2.9) replaced by conditions appropriate to both boundaries being simultaneously free *or* fixed. Their procedure depends on a variational principle and approximate values of S_c^2 are given for the limiting cases $r\to 0$, $r\to\infty$. For comparison with our results we include their values below.

TABLE 1

Values for S_c in isothermal case; after Bdzil and Frisch [2].

	$r\to\infty$		$r\to 0$	
Boundary conditions	a_c	S_c^2	a_c	S_c^2
2 free surfaces	2.2	658	2.0	546
2 rigid surfaces	3.2	1700	2.5	1066

In the above a_c denotes the critical value of the wavenumber, a, at the onset of instability.

For the physically relevant problem (2.8), (2.9) we find it unnecessary to resort to a variational principle: instead we employ an apparently less known technique of Chandrasekhar [4] appropriate to the rigid-free surface problem.

The idea is to shift the fluid domain to $-\frac{1}{2}<z<\frac{1}{2}$ and suppose $(2.9)_1$ holds on $z=-\frac{1}{2}$, but temporarily replace the conditions on the upper plane $z=\frac{1}{2}$ by

$$W=DW=(D^2-a^2)^2(D+r)W=0. \tag{2.10}$$

It follows from the evenness of the operator $(D^2-a^2)^3$ and the boundary conditions which now have to be satisfied at $z=\pm\frac{1}{2}$, that the solutions to (2.8) fall into two noncombining groups of even and odd functions.

The general solution to (2.8) may be expressed as a superposition of solutions of the form $W=e^{\pm qz}$, where q^2 is a root of the equation

$$(q^2-a^2)^3=-S^2a^2. \tag{2.11}$$

By letting $S^2=\tau^3a^4$ the six roots of (2.8) are $\pm iq_0$, $\pm q$, $\pm q^*$, where * denotes complex conjugate and where

$$\begin{aligned}
&q_0=a(\tau-1)^{1/2},\\
&\mathrm{re}(q)=q_1=a\left[\frac{1}{2}(1+\tau+\tau^2)^{1/2}+\frac{1}{2}\left(1+\frac{1}{2}\tau\right)\right]^{1/2},\\
&\mathrm{im}(q)=q_2=a\left[\frac{1}{2}(1+\tau+\tau^2)^{1/2}-\frac{1}{2}\left(1+\frac{1}{2}\tau\right)\right]^{1/2}.
\end{aligned}$$

It will be seen later that it is sufficient to consider only the odd solutions. Then,

$$W=A_0\sin q_0z+A\sinh qz+A^*\sinh q^*z,$$

for constants A_0, A. Boundary conditions $(2.9)_1$, (2.10) then determine three simultaneous linear equations for A_0, A, A^* and the solution of these requires the following determinant to vanish:

$$\begin{vmatrix}
1 & 1 & 1\\
q_0\cot\frac{1}{2}q_0 & q\coth\frac{1}{2}q & q^*\coth\frac{1}{2}q^*\\
q_0\cot\frac{1}{2}q_0+r & \frac{1}{2}(i\sqrt{3}-1)\left(q\coth\frac{1}{2}q+r\right) & -\frac{1}{2}(i\sqrt{3}+1)\left(q^*\coth\frac{1}{2}q^*+r\right)
\end{vmatrix}. \tag{2.12}$$

This condition reduces to

$$\begin{aligned}
&\sqrt{3}\left(q_0\cot\frac{1}{2}q_0+r\right)(q_1\sin q_2-q_2\sinh q_1)(\cosh q_1-\cos q_2)\\
&\quad=(q_1\sinh q_1+q_2\sin q_2)^2-rq_0\cot\frac{1}{2}q_0(\cosh q_1-\cos q_2)^2\\
&\quad\quad+(q_2\sinh q_1-q_1\sin q_2)^2\\
&\quad\quad-(\cosh q_1-\cos q_2)(q_1\sinh q_1+q_2\sin q_2)\left(q_0\cot\frac{1}{2}q_0-r\right).
\end{aligned} \tag{2.13}$$

When the dimensionless reaction rate r (the ratio of the rate of reaction to the rate of diffusion) is large ($r \to \infty$), (2.13) reduces to

$$q_0 \cot \frac{1}{2} q_0 = \frac{(q_1 + q_2\sqrt{3})\sinh q_1 - (q_1\sqrt{3} - q_2)\sin q_2}{\cosh q_1 - \cos q_2}, \tag{2.14}$$

consistent with that obtained by Chandrasekhar [4, p. 41], for the standard rigid-free Bénard problem.

Equation (2.13) is a transcendental equation relating a and τ which we solve numerically. The idea is to determine τ for a given a; the corresponding characteristic value of S^2 then follows. The critical S^2 which governs the nonlinear stability boundary corresponds to the minimum.

The even solutions obtained by taking

$$W = A_0 \cos q_0 z + A \cosh qz + A^* \cosh q^* z,$$

also lead to a characteristic equation for S^2. However, returning to the original problem with

$$W = D^2 W = (D^2 - a^2)^2 W = 0,$$

on the upper surface, it should be noted that the odd solutions satisfy these boundary conditions on $z = 0$. Thus we consider only the odd solutions in the layer between $z = \pm \frac{1}{2}$ and recover the solution to the original problem on a layer of half the depth. The eigenvalue, S^2, to the original problem is the $\frac{1}{16}$th the value obtained from the problem under consideration.

The final results for the case of a lower catalytic rigid boundary and a free upper boundary are given in Table 2.

TABLE 2

r	a_c	S_c^2
10^{10}	2.682	1100.65
100	2.668	1084.64
50	2.653	1070.46
10	2.564	996.50
5	2.492	949.64
2.5	2.412	904.77
0.8	2.304	853.30
0.6	2.285	845.19
0.4	2.264	836.46
0.2	2.241	827.01
0.0	2.215	816.74

Our numerical results are consistent with Chandrasekhar [4] in that when $r \to \infty$ we recover the result of [4], viz. $S_c^2 = 1100.65$.

Notes. 1. It is worth observing that the above method may be employed in the free-free boundary problem considered by Bdzil and Frisch [2].

2. Although the analysis presented here has been restricted to an isothermal system we may, for the nonisothermal case, in the limit $r \to \infty$, deduce the stability boundary easily. In this case (2.6), (2.7) yield

$$(D^2 - a^2)^3 W = -a^2(R^2 + S^2)W,$$

subject to the boundary conditions

$$W = DW = (D^2 - a^2)^2 W = 0, \quad \text{on } z = 0,$$
$$W = D^2 W = (D^2 - a^2)^2 W = 0, \quad \text{on } z = 1.$$

Stability follows in this case provided

$$R^2 + S^2 \leqq 1100.65$$

this result being in agreement with those of Joseph [6], [7].

3. Nonlinear energy stability for case (B). In this section we return to the nonisothermal problem with a thermally insulated lower catalytic plate, the upper surface being stress free. Thus we are investigating the stability of a solution to (1.6) subject to boundary conditions (1.7) case (B). For clarity we rewrite the boundary conditions appropriate to case (B) here,

$$\mathbf{u} = \mathbf{0}, \quad \frac{\partial \phi}{\partial z} = r\phi, \quad \frac{\partial \theta}{\partial z} = -s\phi, \quad z = 0, \tag{3.1}$$

$$\frac{\partial u}{\partial z} = \frac{\partial v}{\partial z} = w = \theta = \phi = 0, \qquad z = 1 \tag{3.2}$$

with $\mathbf{u}, \phi, \theta, p$ still periodic in x, y as stated in the paragraph following (1.7).

We are interested in the case where $H_1 = -1$ and so from (1.4) and (1.5), $H_2 = +1$. Case (B) is mathematically very different from case (A) partly because $H_1 = -1$ and partly due to the boundary conditions on $\partial\theta/\partial z$ at $z = 0$. Both conditions lead to loss of symmetry of the linear time independent operator in (1.6) and so we are no longer able to infer nonlinear stability from an investigation of the equations appropriate to linear stationary convection. Nevertheless, we commence with the energy functional (2.1) and derive the energy equation as

$$\dot{E} = -D(\mathbf{u}) - D(\theta) - D(\phi) + 2S\langle \phi, w \rangle - r\int_\Gamma \phi^2 \, dA + s\int_\Gamma \phi\theta \, dA. \tag{3.3}$$

Because of the integrals over Γ the analysis analogous to that leading from (2.2) to (2.7) yields a complicated Euler–Lagrange system which we would only be able to solve by means of numerical eigenvalue techniques. We have, therefore, developed an alternative technique which avoids heavy numerical work and allows us to use the results of §2 to obtain the nonlinear stability estimates appropriate to this section.

Basically, the idea is to first estimate the integrals over Γ by means of the Cauchy–Schwarz, Poincaré and trace inequalities (see e.g., Bandle [1, p. 101]). As we wish to obtain *quantitative* stability results, however, we need specific values for the constants in these inequalities and so proceed directly as follows. Since $\phi = 0$ on $z = 1$, for x, y fixed we have

$$\phi(x,y,0) = -\int_0^1 \frac{\partial \phi}{\partial z}(x,y,z)\, dz \leqq \left(\int_0^1 \left[\frac{\partial \phi}{\partial z}(x,y,z) \right]^2 dz \right)^{1/2},$$

by Cauchy–Schwarz. Hence, squaring and integrating over Γ,

$$\int_\Gamma \phi^2 \, dA \leqq D(\phi). \tag{3.4}$$

Obviously, the same inequality holds for θ.

Next, the Cauchy–Schwarz and arithmetic-geometric mean inequalities together with (3.4) allow us to show

$$s\int_\Gamma \theta\phi\, dA \leqq \frac{1}{2}s\gamma D(\theta)+\frac{1}{2}s\gamma^{-1}D(\phi), \tag{3.5}$$

where $\gamma=1$. Inequality (3.5) is employed in (3.3) together with the θ-version of (3.4) to yield,

$$\dot{E}\leqq -D(\mathbf{u})+2S\langle\phi,w\rangle-\left(1-\frac{1}{2}s\right)[D(\phi)+D(\theta)]-r\int_\Gamma\phi^2\,dA. \tag{3.6}$$

To use (3.6) we must require

$$s<2. \tag{3.7}$$

Denote by η the number $1-\frac{1}{2}s$ (>0). Then by an argument similar to that used in §2 we obtain from (3.6),

$$\dot{E}\leqq -\mathscr{D}\left(1-\frac{S}{\Lambda}\right)-\eta D(\theta), \tag{3.8}$$

where

$$1|\Lambda=\max\frac{I}{\mathscr{D}}, \tag{3.9}$$

with $I=2\langle\phi,w\rangle$, $\mathscr{D}=D(\mathbf{u})+\eta D(\phi)+A$,[1] and where the maximum is over the set of admissible solutions $\mathbf{u},\phi$. If $S<\Lambda$ then since $\mathscr{D}$ and $D(\theta)$ both satisfy Poincaré's inequality we may derive an inequality of the form $\dot{E}\leqq-\mu E$, $\mu>0$, from which nonlinear stability follows.

The nonlinear stability problem is, therefore, reduced to solving the maximum problem (3.9). We are led naturally to the Euler–Lagrange equations[2]

$$\Delta\mathbf{u}+S_E\phi\mathbf{k}-\nabla p=0,\qquad S_E w+\eta\Delta\phi=0, \tag{3.10}$$

to be solved subject to boundary conditions (3.1), (3.2). In (3.10) S_E is an eigenvalue which arises from the introduction of a Lagrange multiplier. It is easy to show that $S_E=\Lambda$ and so $S<S_E$ is a sufficient condition to guarantee the decay of not necessarily infinitesimal disturbances. The functions u,v,p,ϕ are next eliminated from (3.10) and then due to linearity and the periodicity of w we may write $w=W(z)\cdot\exp[i(kx+my)]$ to obtain,

$$(D^2-a^2)^3W=-\eta^{-1}S_E^2a^2W,\qquad z\in(0,1), \tag{3.11}$$

the solution to which must satisfy,

$$\begin{aligned} &W=D^2W=(D^2-a^2)^2W=0, && z=1,\\ &W=DW=(D^2-a^2)^2(D-q)W=0, && z=0. \end{aligned} \tag{3.12}$$

---

[1] $A=r\int_\Gamma\phi^2\,dA$.

[2] With $q=r/\eta$; cf. [9].

Of course, (3.11) and (3.12) may be identified with (2.8) and (2.9) and so we may use the results in Table 2 to deduce nonlinear stability in case (B), provided also the restriction (3.7) is satisfied. To illustrate, we give two examples. The equilibrium solution (1.3) is nonlinearly stable provided

$$r\to 0, \quad s<2, \quad S^2<816.74\left(1-\frac{1}{2}s\right), \tag{3.13}$$

or

$$r=0.6\left(1-\frac{1}{2}s\right), \quad s<2, \quad S^2<845.19\left(1-\frac{1}{2}s\right). \tag{3.14}$$

It might be observed that our stability criterion places no restriction on the size of the Rayleigh number. This is due to the fact that $H_1=-1$; in the absence of reaction this corresponds to heating the layer from above and a similar conclusion is possible there, see Joseph [7].

In the limiting case $r\to 0$, $s\to 0$ both the linear and energy stability boundaries are easily obtained and serve a very useful purpose. Assuming stationary convection the linear equations which arise from (1.6) yield

$$(D^2-a^2)^3W=-a^2(S^2-R^2)W,$$

with the boundary conditions

$$\begin{aligned} w=D^2W=(D^2-a^2)^2W=0, &\qquad \text{on } z=1,\\ W=DW=D(D^2-a^2)^2W=0, &\qquad \text{on } z=0. \end{aligned}$$

Again the results of Table 2 apply and so the instability boundary is given by

$$S^2-R^2=816.74.$$

From (3.13) there is nonlinear energy stabilty provided

$$S^2<816.74.$$

Therefore, there is the possibility of oscillatory convection (subcritical bifurcation) for S^2 in the region

$$816.74<S^2<816.74+R^2.$$

Acknowledgment. Carol McTaggart is deeply grateful to Professor I. N. Sneddon for helpful advice and discussions.

REFERENCES

[1] C. BANDLE, *Isoperimetric Inequalities and Applications*, Pitman, London, 1980.

[2] J. BDZIL AND H. L. FRISCH, *Chemical instabilities. II. Chemical surface reactions and hydrodynamic instability*, Phys. Fluids, 14 (1971), pp. 475–481.

[3] _____, *Chemical instabilities. IV. Nonisothermal chemical surface reactions and hydrodynamic instability*, Phys. Fluids, 14 (1971), pp. 1077–1086.

[4] S. CHANDRASEKHAR, *Hydrodynamic and Hydromagnetic Stability*, University Press, Oxford, 1961.
[5] G. P. GALDI AND B. STRAUGHAN, *Exchange of stabilities, symmetry and nonlinear stability*, Arch. Rational Mech. Anal., to appear.
[6] D. D. JOSEPH, *Stability of Fluid Motions*, *Vol.* II, Springer-Verlag, Berlin-Heidelberg-New York, 1976.
[7] ———, *Global stability of the conduction-diffusion solution*, Arch. Rational Mech. Anal., 36 (1970), pp. 285–292.
[8] D. E. LOPER AND P. H. ROBERTS, *On the motion of an iron-alloy core containing a slurry*, Geophys. Astrophys. Fluid Dynamics, 9 (1978), pp. 289–321.
[9] G. P. GALDI AND B. STRAUGHAN, *Convection in thawing subsea permafrost*, to appear.

SIAM J. MATH. ANAL.
Vol. 17, No. 2, March 1986

006

ON THE SPECTRAL PROPERTIES OF A CLASS OF ELLIPTIC FUNCTIONAL DIFFERENTIAL OPERATORS ARISING IN FEEDBACK CONTROL THEORY FOR DIFFUSION PROCESSES*

A. VAN HARTEN†

Abstract. In this paper Dirichlet boundary value problems are considered for certain operators of the form $L+\Pi$, where L is a 2nd order, elliptic, formally self-adjoint *PDO* and Π is a feedback operator with finite-dimensional range.

The results concern mainly the resolvent $(L+\Pi-\lambda)^{-1}$, the analyticity of the semigroup generated by $L+\Pi$, the location of the spectrum $\sigma(L+\Pi)$ and the completeness of the eigenspaces and the eigenprojections associated to $\sigma(L+\Pi)$.

As a consequence of the completeness of the eigenprojections it is possible to derive quite a number of interesting formulas of the "resolution of the identity" type.

1. Introduction. In this paper we shall consider problems of Dirichlet type on a bounded domain $D\subset\mathbb{R}^d$ for a class of operators of the form $L+\Pi$. Here L will be a linear, uniformly elliptic, formally selfadjoint, 2nd order partial differential operator with time-independent real coefficients

$$L=\sum_{i=1}^{d}\sum_{j=1}^{d}\frac{\partial}{\partial x_i}\left(a_{ij}\frac{\partial}{\partial x_j}\right)+a_0. \tag{1.1}$$

For the sake of simplicity we shall usually assume that the domain D has a smooth boundary ∂D and that the coefficients of L are elements of $C^\infty(\overline{D})$.

Of course $\overline{D}$ denotes the closure of D in $\mathbb{R}^d$.

The matrix $A=(a_{ij})$ is strictly positive definite, uniformly on $\overline{D}$ and symmetric. The structure of Π will be as follows:

$$\Pi=\sum_{i=1}^{p}c_iP_i,\qquad p<\infty. \tag{1.2}$$

The c_i's will be real functions only dependent of the space-coordinates $x\in\overline{D}$. They will be chosen as elements of some Banach space Y of real functions on $\overline{D}$.

The P_i's will be time-independent real continuous linear functionals on some Banach space X of real functions on $\overline{D}$, i.e. the P_i's are elements of $X'=$ the dual space of X.

Therefore the operator Π can be interpreted as an element of $\mathscr{L}(X\to Y)=$ the space of continuous linear operators from X into Y. Due to the requirement $p<\infty$ the operator Π has a finite-dimensional range, i.e. is degenerate (see Kato (1966)). A striking difference between the operator Π and the differential operator L is that Π possesses a nonlocal character. This explains that $L+\Pi$ is referred to as a functional differential operator in the title of this paper.

*Received by the editors July 2, 1980, and in final revised form September 12, 1984.
†Rijksuniversiteit Utrecht, Mathematisch Institut, 3508 TA Utrecht, the Netherlands.

The problems we shall look at can now be formulated as follows:

$$\frac{\partial v}{\partial t} = (L+\Pi)v, \tag{1.3}$$

$v=0$ on ∂D, boundary conditions of Dirichlet type;

$v(\cdot,0)=\Psi$, initial conditions,

$$(L+\Pi-\lambda)u=g, \tag{1.4}$$

$u=0$ on ∂D, boundary conditions of Dirichlet type.

Note that (1.3) constitutes a set of equations for the evolution in time of the function v, whereas (1.4) is a stationary problem for the function u.

In (1.4) λ has to be considered as a spectral parameter and the function g is some inhomogeneous term.

The emphasis in this paper will be on the spectral analysis associated to (1.4).

1.1. On the relation to feedback control problems. The topics considered in this paper have their origin in the study of controlled heat-diffusion processes. The control applied to such a process will be instantaneous and of an automatic feedback type. The heating/cooling of this feedback mechanism will flow directly into the domain (i.e. not through the boundary). If we further suppose that no convection of heat will take place and that the temperature is given on the boundary ∂D, then a homogeneous version of a mathematical model which describes the evolution in time of the temperature is given by (1.3). In (1.3) the effect of diffusion of heat and direct exchange with the surroundings is incorporated in L. The effect of the feedback control is described by the operator Π. In this context the functions c_i ar called the "control inputs" and the P_i's are called "observers". The operator Π will sometimes be referred to as the feedback control operator.

Very simple illustrations of what we have in mind may serve. The following examples have $p=1$, $\Pi=c_1P_1$.

(i) $c_1\in H^s(D)$ for some $s\geqq 0$, $P_1\in L_2(D)'$ f.e. the observer P_1 observes the average over D: $P_1w=\int_D w(x)\,dx$ for $w\in L_2(D)$;

(ii) $c_1\in C^\alpha(\overline{D})$ for some $\alpha\geqq 0$, $P_1\in C(D)'$ f.e. the observer P_1 observes the value in a point $y\in\overline{D}$: $P_1w=w(y)$ for $w\in C(\overline{D})$.

Here the following notation is used:

$L_2(D)$ = space of equivalence classes of square integrable functions on D;

$H^s(D)$ = Sobolev space of order S on D; (see Adams (1975); Lions and Magenes (1971));

$C^\alpha(\overline{D})$ = Hölder space of order α on $\overline{D}$ (see Adams, (1975); Ladyzhenskaya and Ural'tseva (1968)).

Other examples are explored in van Harten, Schumacher (1980). For further information on this type of control problems we refer to Curtain and Pritchard (1978). For a related class of problems in which the control enters in the b.c. we refer to Triggiani (1980) and references given there.

1.2. Description of the contents. The results of this paper can be divided into two groups. The first group of results concerns the solvability theory for the problem (1.4), the properties of the resolvent $(L+\Pi-\lambda)^{-1}$ and the solution of (1.3) in terms of the analytic semigroup generated by $L+\Pi$. These results are found in the §§3 and 4 and they can be considered as basic material for the second group of results. The results on the analytic semigroup generated by $L+\Pi$ are a generalization of an analogous result

for bounded Π, see Curtain and Pritchard (1978, Thm. 2.31) and also of Kato (1966, pp. 498–499). The second group contains results on the location of the spectrum associated to (1.4) and on the completeness of the eigenspaces and eigenprojections associated to this spectrum. The completeness results given here are certainly nontrivial since in general $L+\Pi$ is not formally self-adjoint in any sense. The results on completeness of eigenprojections are generalizations of Kato (1966, Thms. 4.15 and 4.16, Chap. V, §4.5). From these results we can deduce very interesting "resolution of the identity" formula's in $L_2(D)$ and certain other Hilbert spaces. These "resolutions of the identity" look promising with respect to applications, also applications outside the control context. The second group of results can be found in the Theorems 5, 6, 7 and 8.

Let us conclude the introduction with a few remarks. Our first remark concerns the spaces X and Y mentioned in the description of the operator Π.

It is important to notice that for a given problem it is always possible to choose the spaces X and Y in several ways. It is then of course logical to take Y as "small" as possible and X as "large" as possible in order to get a most significant theory. In any way building a theory for these problems (1.3)–(1.4) will depend strongly on what one takes for X and Y. The choices which we allow for X and Y will be specified further on.

Secondly, though from a physical point of view it is natural to look at L and Π as operators on real functions, it will sometimes be profitable to extend L and Π to complex functions. Notationally we shall not distinguish between real operators and function spaces and their obvious complex extensions.

2. The uncontrolled problem. The main purpose of this section is to introduce some notations and concepts, which will be used further on. First this will be done for the solution(s), their regularity properties, the spectrum and the eigenprojections associated with the problem:

$$(L-\lambda)u=g, \tag{2.1}$$

$$u=0 \quad \text{on } \partial D \quad (BC). \tag{2.2}$$

In §2.2 we introduce some notation concerning fractional powers of $-L+\lambda_0$, $\lambda_0 \in \mathbb{R}$ sufficiently large.

2.1. Solutions of (2.1), (2.2); spectrum and eigenprojections. It is well known that for $\lambda=\lambda_0$, with λ_0 real and sufficiently large >0 the resolvent $(L-\lambda_0)^{-1}$ is a compact, selfadjoint operator from $L_2(D)$ into $L_2(D)$ (see Trèves (1975)). Consequently the spectrum $\sigma(L)$ associated to (2.1)–(2.2) on $L_2(D)$ consists of a denumerable infinite set of isolated eigenvalues on the real axis bounded at the positive side and extending towards $-\infty$ without accumulation points (see Dunford and Schwartz (1963)). Each eigenvalue has a finite algebraic multiplicity and the number of linearly independent eigenfunctions corresponding with an eigenvalue equals its algebraic multiplicity. Further the eigenfunctions can be chosen in such a way that they form an orthonormal, complete basic system in $L_2(D)$.

For the sequence of eigenvalues and corresponding linearly independent orthonormal eigenfunctions we shall use the notation:

$$\mu_n, \phi_n \quad \text{with } n \in N. \tag{2.1.1}$$

The numbering is such that

$$k>n \quad \Rightarrow \quad \mu_k \leqq \mu_n \text{ and } \langle \phi_k, \phi_n \rangle_{L_2(D)} = \delta_{n,k}. \tag{2.1.2}$$

In this numbering each eigenvalue is repeated according to its algebraic multiplicity.

For the system $\{\phi_n; n\in\mathbb{N}\}$ we have the "resolution of the identity" formula inherent to the completeness of this system in $L_2(D)$:

$$I=\sum_{n=1}^{\infty}\phi_n\langle\phi_n|\cdot\rangle_{L_2(D)}. \tag{2.1.3}$$

For $\lambda\in\sigma(L)$ we shall denote by $E(\lambda)$: $L_2(D)\to L_2(D)$ the orthogonal projection on the eigenspace corresponding to λ. If $\lambda\in\sigma(L)$ we define $J(\lambda)=\{n\in\mathbb{N}\,|\,\mu_n=\lambda\}$. Hence

$$E(\lambda)=\sum_{n\in J(\lambda)}\phi_n\langle\phi_n|\cdot\rangle_{L_2(D)}. \tag{2.1.4}$$

To conclude this subsection we mention some rather elementary facts on the growth of the eigenvalues for $n\to\infty$.

Using Minakshisundaram and Pleyel (1949) or Garabedian (1964) we see that for n sufficiently large the eigenvalues satisfy

$$R_1 n^{\nu}\leqq|\mu_n|\leqq R_2 n^{\nu},\qquad n\geqq n_0. \tag{2.1.5}$$

Here R_1, R_2 are real constants >0 and $\nu=2/d$ with d the dimension of D.

The problem (2.1)–(2.2) possesses a unique solution for all $g\in L_2(D)$ if and only if $\lambda\notin\sigma(L)$. This solution will be denoted by

$$u=(L-\lambda)^{-1}g.$$

In the case $\lambda\in\sigma(L)$ the problem (2.1)–(2.2) possesses a solution if and only if $g\in L_2(D)$ satisfies $E(\lambda)g=0$. Under this condition the solution space of (2.1)–(2.2) is given by

$$\begin{aligned}&u\in(L-\lambda)^{-1}g\Leftrightarrow\\&u=(L-\lambda)_*^{-1}g-u_1\quad\text{with}\\&(L-\lambda)_*^{-1}=\sum_{n\notin J(\lambda)}\phi_n(\mu_n-\lambda)^{-1}\langle\phi_n,g\rangle_{L_2(D)},\\&u_1\in\operatorname{span}\{\phi_n|n\in J(\lambda)\}.\end{aligned} \tag{2.1.6}$$

The operator $(L-\lambda)_*^{-1}$ will be called the restricted resolvent at $\lambda\in\sigma(L)$. Once a solution of (2.1)–(2.2) exists, it possesses the following regularity properties:

$$\begin{aligned}&g\in H^s(D)\Rightarrow u\in H^{s+2}(D),\quad s\geqq 0,\quad s\notin\mathbb{N}-\tfrac{1}{2},\\&g\in C^{\alpha}(\overline{D})\Rightarrow u\in C^{\alpha+2}(\overline{D}),\quad \alpha>0,\quad \alpha\notin\mathbb{N}.\end{aligned} \tag{2.1.7}$$

Moreover the operators $(L-\lambda)^{-1}$, $\lambda\notin\sigma(L)$ and $(L-\lambda)_*^{-1}$, $\lambda\in\sigma(L)$ are bounded in each sense suggested by (2.1.7). For these results we refer to: Lions and Magenes (1972, Part I) and Ladyzhenskaya and Ural'tseva (1968).

Note that as a consequence of (2.1.7) all eigenfunctions are elements of $C^{\infty}(\overline{D})$.

2.2. Fractional powers of $-L+\lambda_0$, $\lambda_0\in\mathbb{R}$ sufficiently large. In this section we shall suppose that $\lambda_0\in\mathbb{R}$ satisfies

$$\lambda_0\geqq\max_{\overline{D}}a_0+1\underset{\text{def}}{=}\mu_0. \tag{2.2.1}$$

Let us introduce the simplified notation A for the operator induced by $-L+\lambda_0$ with Dirichlet boundary conditions on either $L_2(D)$ or $C(\overline{D})$. Using variational techniques and the maximum principle for 2nd order elliptic P.D.E., it is easy to check that the

resolvent $(s+A)^{-1}$ with $s\in\mathbb{R}$, $s>0$ satisfies the following estimates:

$$\|\|\|(s+A)^{-1}\|\|\|_{L_2(D)}\leqq(1+s)^{-1}, \tag{2.2.2}$$

$$\|\|\|(s+A)^{-1}\|\|\|_{C(\overline{D})}\leqq(1+s)^{-1}. \tag{2.2.3}$$

Here we use the notation $\|\|\|\ \|\|\|_X$ for the usual norm on the space of bounded operators from the Banach space X into itself.

The operator A defined on $L_2(D)$ is closed, its domain is dense is $L_2(D)$ and because of (2.3.2) this operator is of positive type (see Krasnoselskii (1976, p. 279)).

Of course it is not true that the operator A defined on $C(\overline{D})$ has a domain dense in $C(\overline{D})$. This is remedied by restricting the operator to

$$C_0(\overline{D})=\{f\in C(\overline{D})\,|\,f=0 \text{ on } \partial D\}.$$

On $C_0(\overline{D})$ the operator A is also closed, densely defined and of positive type.

These facts induce that on $L_2(D)$ or $C_0(\overline{D})$ fractional powers of A are well defined; see Krasnoselskii (1976, pp. 280–288).

It will be very useful to introduce some Hilbert and Banach spaces connected to these fractional powers of A. For $\alpha\geqq 0$ we define:

$$HD(\alpha)=\text{range of } A^{-\alpha} \text{ on } L_2(D)=\text{domain of } A^{\alpha} \text{ w.r.t. } L_2(D), \tag{2.2.4}$$
$$\|u\|_{HD(\alpha)}=\|A^{\alpha}u\|_{L_2(D)},\quad \langle u,v\rangle_{HD(\alpha)}=\langle A^{\alpha}u,A^{\alpha}v\rangle_{L_2(D)},$$

$$C_0D(\alpha)=\text{range of } A^{-\alpha} \text{ on } C_0(\overline{D})=\text{domain of } A^{\alpha} \text{ w.r.t. } C_0(\overline{D}), \tag{2.2.5}$$
$$\|u\|_{C_0D(\alpha)}=\|A^{\alpha}u\|_{C(\overline{D})},$$

$$CD(\alpha)=\{u\in HD(\alpha)\,|\,A^{\alpha}f\in C(\overline{D})\}, \tag{2.2.6}$$
$$\|u\|_{CD(\alpha)}=\|A^{\alpha}u\|_{C(\overline{D})}.$$

It is not difficult to check that different choices of λ_0 lead to equivalent spaces and norms, see also van Harten (1979).

Note that for $\alpha_1>\alpha_2$ we have $HD(\alpha_1)\subset_{\text{dense}} HD(\alpha_2)$, $C_0D(\alpha_1)\subset_{\text{dense}} C_0D(\alpha_2)$, see Krasnoselskii (1976, pp. 286). It is also true that $\alpha_1>\alpha_2\Rightarrow CD(\alpha_1)\subset CD(\alpha_2)$, but in this case the inclusion is not necessarily dense (see Appendix, Proposition A.4).

The spaces $HD(\alpha)$, $CD(\alpha)$ and $C_0D(\alpha)$ will play a dominant role in the theory to be developed further on.

A very natural and important question about the spaces and norms introduced in (2.2.4)–(2.2.6) is how they are related to the Sobolev and Hölder spaces, $H^s(D)$ and $C^s(\overline{D})$, $s>0$. It is possible to show, that

$$HD(\alpha)\cong\{u\in H^{2\alpha}(D)\,|\,L^k u=0 \text{ on } \partial D,\ k \text{ integer},\ 0\leqq 2k<2\alpha-\tfrac{1}{2}\}, \tag{2.2.7}$$
if $2\alpha+\frac{3}{2}\in\mathbb{N}$;

$$C_0D(\alpha)\subset\{u\in C^{2\beta}(\overline{D})\,|\,L^k u=0 \text{ on } \partial\partial,\ k \text{ integer},\ 0\leqq k\leqq\beta\}, \text{ if } 0\leqq\beta<\alpha; \tag{2.2.8}$$
$$C_0D(\alpha)\supset\{u\in C^{2\beta}(\overline{D})\,|\,L^k u=0 \text{ on } \partial\partial,\ k \text{ integer},\ 0\leqq k\leqq\alpha\}, \text{ if } \beta>\alpha;$$

$$A^{-\beta}CD(\alpha)\supset C_0D(\alpha+\beta-\varepsilon),\ \alpha\geqq 0,\ \beta>0,\ 0\leqq\varepsilon<\alpha+\beta \tag{2.2.9}$$
or equivalently,
$$CD(\alpha)\subset C_0D(\alpha-\varepsilon),\qquad 0<\varepsilon<\alpha,$$
$$CD(\alpha)\supset C_0D(\alpha).$$

All these inclusions are accompanied by continuous injections.

The relation between $HD(\alpha)$ and Sobolev spaces as given in (2.2.7) is well known (see for example Lions and Magenes (1972, Part I, Remark 2.3, p. 10)).

To this author's knowledge the contents of (2.2.8)–(2.2.9) though not difficult to prove, are not available in the literature. The proofs of the statements (2.2.8) and (2.2.9) can be found in an Appendix.

Let us now look at the problem (2.1)–(2.2) with g in the space $HD(\alpha)$, $CD(\alpha)$ or $C_0D(\alpha)$. The list given in (2.1.7) for the regularity of a solution can be extended as follows:

$$
\begin{aligned}
&g\in HD(\alpha)\Rightarrow u\in HD(\alpha+1), && \alpha\geqq 0,\\
&g\in CD(\alpha)\Rightarrow u\in CD(\alpha+1), && \alpha\geqq 0,\\
&g\in C_0D(\alpha)\Rightarrow u\in C_0D(\alpha+1), && \alpha\geqq 0.
\end{aligned}
\tag{2.2.10}
$$

This is easily seen by rewriting (2.1)–(2.2) as $Au=(\lambda_0-\lambda)u-g$.

We conclude this section with a remark on the eigenfunctions $\{\phi_n\colon n\in\mathbb{N}\}$ introduced in §2.1.It is straightforward to check, that $\{\phi_n/\|\phi_n\|_{HD(\alpha)}; n\in\mathbb{N}\}$ defines an orthonormal sequence in $HD(\alpha)$. Using the selfadjointness of A^{-1} on $HD(\alpha)$, we see that this orthonormal sequence of functions is also complete in $HD(\alpha)$. As a consequence analogues of (2.1.3) and (2.1.4) can be given in $HD(\alpha)$.

3. The controlled problem: application of the Weinstein–Aronszajn theory. For future reference we shall now present some well-known basic material on the solvability of the problem:

$$(L+\Pi-\lambda)u=g, \tag{3.1}$$

$$u=0 \quad \text{on } \partial D, \quad \text{boundary condition of Dirichlet type.} \tag{3.2}$$

This theory is essentially due to Weinstein and Aronszajn and it is built in an essential way on the degeneracy, i.e. finite-dimensional range property, of the operator $\Pi=\sum_{i=1}^{p}c_iP_i$ with $c_i\in Y$ and $P_i\in X'$.

For the Banach spaces X, Y we allow the following choices:

$$
\begin{aligned}
&\text{a.}\quad X=H^{s+2}(D),\ Y=H^{s}(D), && s\geqq 0,\ s\notin -\tfrac{1}{2}+N,\\
&\text{b.}\quad X=C^{\alpha+2}(\overline{D}),\ Y=C^{\alpha}(\overline{D}), && \alpha\geqq 0,\ \alpha\notin\mathbb{N} \quad \text{or}\\
&\text{c.}\quad X=HD(\alpha+1),\ Y=HD(\alpha), && \alpha\geqq 0 \quad \text{or}\\
&\text{d.}\quad X=CD(\alpha+1),\ Y=CD(\alpha), && \alpha\geqq 0 \quad \text{or}\\
&\text{e.}\quad X=C_0D(\alpha+1),\ Y=C_0D(\alpha), && \alpha\geqq 0.
\end{aligned}
\tag{3.3}
$$

In our consideration we shall always take

$$g\in Y. \tag{3.4}$$

Next we shall look for a solution u of (3.1)—(3.2) in the space X. Note, that L is a closed operator on Y with domain X and that because of (2.1.7), (2.2.10) Π is relatively bounded w.r.t L. As a consequence the theory as given in Kato (1966, Chap. IV, §6) can be applied and it leads to the following results.

First we consider case (i), where $\lambda\notin\sigma(L)$. Let c denote the p-vector of control inputs, i.e. $c\in Y^P$. Analogously we define p as the p-vector of observers i.e. $P\in\{X'\}^P$.

Let $\Omega(\lambda)$ be the $p\times p$ matrix with the following matrix elements:

$$[\Omega(\lambda)]_{i,j}=P_i(L-\lambda)^{-1}c_j+\delta_{ij}. \tag{3.5}$$

In case (i) the problem (3.1), (3.2) is uniquely solvable if and only if the matrix $\Omega(\lambda)$ is nonsingular.

The solution is then given by

$$(3.6)\quad u=(L+\Pi-\lambda)^{-1}g=(L-\lambda)^{-1}g-\left\langle (L-\lambda)^{-1}c, \Omega(\lambda)^{-1}P(L-\lambda)^{-1}g\right\rangle,$$

where $\langle\ ,\ \rangle$ denotes the obvious pairing between X^P and $\mathbb{C}^P$. The resolvent $(L+\Pi-\lambda)^{-1}$ is an element of $\mathscr{L}(Y\to X)$.

If for $\lambda\notin\sigma(L)$ it holds true that the matrix $\Omega(\lambda)$ is singular then λ is a point of the spectrum associated to to (3.1), (3.2). The homogeneous problem then possesses nontrivial solutions which are given by

$$(3.7)\qquad u=\left\langle (L-\lambda)^{-1}c,\xi\right\rangle \text{ with } \xi\in\text{null space of } \Omega(\lambda).$$

Let us now consider case (ii), where $\lambda\in\sigma(L)$. We introduce the notation $J(\lambda)=\{k+i\,|\,1\leqq i\leqq m\}$ with m the algebraic multiplicity of the eigenvalue $\lambda\in\sigma(L)$, and we define terms as follows. Let $\phi\in X^m$ be the vector of eigenfunctions $\phi_{k+1},\cdots,\phi_{k+m}$. By ϕ' we shall denote the element of $\{Y'\}^m$ with components

$$\langle\phi_{k+1},\cdot\rangle_{L_2(D)},\cdots,\langle\phi_{k+m},\cdot\rangle_{L_2(D)}.$$

Let us introduce the following matrices

$$\begin{aligned}
&[\Omega_*(\lambda)]_{i,j}=P_i(L-\lambda)_*^{-1}c_j+\delta_{ij}, && \text{a } p\times p \text{ matrix},\\
&[P\phi^T]_{i,j}=P_i\phi_{k+j}, && \text{a } p\times p \text{ matrix},\\
&[\phi' c^T]_{i,j}=\langle\phi_{k+i},c_j\rangle_{L_2(D)}, && \text{a } m\times p \text{ matrix}.
\end{aligned}$$

Further we need the following $(p+m)\times(p+m)$ matrix:

$$(3.8)\qquad \hat{\Omega}(\lambda)=\begin{pmatrix}\Omega_*(\lambda) & P\phi^T\\ \phi' c^T & 0\end{pmatrix}.$$

It is easy to see that in case (ii) the problem (3.1.2) is uniquely solvable if and only if the matrix $\hat{\Omega}(\lambda)$ is nonsingular. The solution is then given by

(3.9)

$$u\underset{\text{def}}{=}(L+\Pi-\lambda)_*^{-1}g=(L-\lambda)^{-1}g+\left\langle\begin{pmatrix}(L-\lambda)_*^{-1}c\\ \phi\end{pmatrix},\quad \hat{\Omega}(\lambda)^{-1}\begin{pmatrix}P(L-\lambda)_*^{-1}\\ \phi'\end{pmatrix}g\right\rangle.$$

In these circumstances $(L+\Pi-\lambda)^{-1}$ is again an operator $\in\mathscr{L}(Y\to X)$. If for $\lambda\in\sigma(L)$ it holds true that the matrix $\hat{\Omega}(\lambda)$ is singular then λ is a point of the spectrum associated with (3.1–2).

The homogeneous problem then possesses nontrivial solutions given by

$$(3.10)\qquad u=\left\langle\begin{pmatrix}(L-\lambda)_*^{-1}c\\ \phi\end{pmatrix},\quad \begin{pmatrix}\xi\\ \zeta\end{pmatrix}\right\rangle \text{ with } \begin{pmatrix}\xi\\ \zeta\end{pmatrix}\in\text{null space of } \hat{\Omega}(\lambda).$$

Hence, let $\sigma(L+\Pi)$ be the spectrum associated with the controlled operator. Then a Fredholm alternative is valid: $\lambda\in\sigma(L+\Pi)\Leftrightarrow$ the problem (3.1)–(3.2) possesses non-trivial solutions in the space X. Moreover,

(3.11) $\lambda\in(L+\Pi)\Leftrightarrow$ (i) $\lambda\notin\sigma(L)$ and the matrix $\Omega(\lambda)$ (see (3.7)) is singular or
(ii) $\lambda\in\sigma(L)$ and the matrix $\dot{\Omega}(\lambda)$ (see (3.10)) is singular.

If $\lambda\notin\sigma(L+\Pi)$ then the resolvent of (3.1)–(3.2) $(L+\Pi-\lambda)^{-1}$ is a compact operator from Y into Y, because of the compact imbedding of x into y. Hence for the spectrum $\sigma(L+\Pi)$ there are two possibilities.

1. The normal case: $\sigma(L+\Pi)$ consists of a denumerable sequence of eigenvalues of finite algebraic multiplicity without accumulation points,

2. the super-singular case: $\sigma(L+\Pi)=\mathbb{C}$.

In §7 an example will be given where the spectrum associated to (3.1)–(3.2) is super-singular.

It is also possible to describe $\sigma(L+\Pi)$ in terms of the following meromorphic function:

(3.12) $$\omega(\lambda)=\det\Omega(\lambda) \quad \text{with } \Omega(\lambda) \text{ as in (3.5)}.$$

It appears that $\omega(\lambda)$ is the so-called Weinstein–Aronszajn determinant of the operator $L+\Pi$ with Dirichlet b.c. (see Kato (1966, Chap. IV, §6)). The following results are a direct consequence of the theory given there.

Let $i(\lambda_0;\omega)$ denote the Laurent index of the meromorphic function $\omega(\lambda)$ for $\lambda\to\lambda_0$, i.e. if $\omega\not\equiv 0$ then $i(\lambda_0;\omega)$ is the number $i\in\mathbb{Z}$ such that $\omega(\lambda).(\lambda-\lambda_0)^{-i}$ has a finite limit $\neq 0$ for $\lambda\to\lambda_0$; if $\omega\equiv 0$ then $i(\lambda_0;\omega)=\infty$ for all $\lambda_0\in\mathbb{C}$. Let $m(\lambda_0;L)$ denote the multiplicity of λ_0 with respect to $\sigma(L)$, i.e. if $\lambda_0\notin\sigma(L)$ then $m(\lambda_0;L)=0$; if $\lambda_0\in\sigma(L)$ then $m(\lambda_0;L)=\sum J(\lambda_0)$ with $J(\lambda_0)$ as defined in (2.1.4). Now the following characterization is valid:

(3.13) $$\lambda_0\in\sigma(L+\Pi)\Leftrightarrow i(\lambda_0;\omega)+m(\lambda_0;L)>0.$$

If $\sigma(L+\Pi)\neq\mathbb{C}$ then for each $\lambda_0\in\sigma(L+\Pi)$ its algebraic multiplicity is given by $i(\lambda_0;\omega)+m(\lambda_0,\omega)$. This result shows clearly that each point of $\sigma(L+\Pi)$ is either in $\sigma(L)$ or is a zero of the function $\omega(\lambda)$.

The solvability theory and the propertries of the spectrum will frequently be used in the sequel.

4. Exponential stability of the null-solution of the time-dependent controlled problem. It is well known that in the uncontrolled case the operator L with Dirichlet boundary conditions generates an analytic semigroup on each of the spaces $HD(\alpha)$, $C_0D(\alpha)$, $\alpha\geqq 0$ (see Dunford and Schwartz (1963) and Stewart (1974), (1980), respectively). Furthermore the stability of the null-solution can be characterized in terms of $\nu=\sup_{\lambda\in\sigma(L)}\mathrm{Re}\lambda$. Analogous results hold in the case of the controlled system, see Theorem 4.1. If the control operator Π is a bounded operator from $HD(\beta)$ into $HD(\alpha)$ as in case a from Theorem 4.1 or from $C_0D(\beta)$ into $C_0D(\alpha)$ with $0\leqq\alpha\leqq\beta<\alpha+1$ this is a direct consequence of a perturbation result for analytic semigroups see Kato (1966, pp. 497–498)). This result is applicable here, because Π is an A (see §2.3) bounded operator with A-bound equal to zero, as one can easily verify using interpolation inequalities.

However, in the Banach space case we can improve on this result. We shall show that $L+\Pi$ generates an analytic semigroup on $C_0D(\alpha)$ even if $\Pi=\sum_{i=1}^{p}c_iP_i$ is a bounded map from $C_0D(\beta)$ into $CD(\alpha)$ with $0\leqq\alpha\leqq\beta<\alpha+1$. In comparison with the previous result where range $\Pi\subset C_0D(\alpha)$ we can now allow for a more general class of control functions c_i: $i=1,\cdots,p$. For example, if $\alpha=0$ it is no longer necessary that $c_i=0$ on the boundary ∂D. It will be clear that Π is no longer relatively bounded w.r.t. A in this situation. This improved result is formulated in Theorem 4.1(b). The proof of that result is such that "mutatis mutandis" it also covers case (a).

THEOREM 4.1. *Let* Π *be* $\in\mathscr{L}(X\to Y)$ *with*

$$\begin{aligned}&\text{a.}\quad X=HD(\beta),\quad Y=HD(\alpha),\quad 0\leqq\alpha\leqq\beta<\alpha+1\quad \textit{or}\\ &\text{b.}\quad X=C_0D(\beta),\quad Y=CD(\alpha),\quad 0\leqq\alpha\leqq\beta<\alpha+1.\end{aligned}\tag{4.1}$$

Then $L+\Pi$ *with Dirichlet boundary conditions generates an analytic semigroup on the space* Y_0 *with*

$$\begin{aligned}&\text{a.}\quad Y_0=HD(\alpha),\\ &\text{b.}\quad Y_0=C_0D(\alpha).\end{aligned}\tag{4.2}$$

The solution of (1.3): $v(\cdot,t)=e^{(L+\Pi)t}\Psi$ *satisfies the usual estimate for* $t\to\infty$

$$\|v(\cdot,t)\|_{Y_0}\leqq K(\varepsilon)e^{(\mu+\varepsilon)t}\|\Psi\|_{Y_0}\quad\forall\Psi\in Y_0\tag{4.3}$$

with $\mu=\sup_{\lambda\in\sigma(L+\Pi)}\operatorname{Re}\lambda$, $\varepsilon>0$ *arbitrarily small. If* $\mu<0$ *this implies exponential stability of the null solution of* (1.1)–(1.3) *with respect to initial perturbations in* Y_0.

Proof of Theorem 4.1. We consider case (b). We shall first show that $\exists\sigma_0>0$ such that $\forall\lambda$ with $\operatorname{Re}\lambda\geqq\sigma_0$:

$$\left\|\left|(L+\Pi-\lambda)^{-1}\right|\right\|_{Y_0}\leqq K(1+|\lambda|)^{-1}.\tag{4.4}$$

In order to derive (4.4) we make some preparations.

Step 1. $\exists\sigma_0\geqq 0\ \exists K>0\ \forall\lambda$ with $\operatorname{Re}\lambda\geqq\sigma_0\ \forall f\in C(\bar{D})$:

$$\|(L-\lambda)^{-1}f\|\leqq\frac{K}{1+|\lambda|}\|f\|\tag{4.5}$$

with $\|\cdot\|=\|\cdot\|_{C(\bar{D})}$.For $f\in C_0(\bar{D})$ this estimate is a consequence of the fact that L generates a semigroup on $C_0(\bar{D})$ (see Krasnoselskii (1976, pp. 270)). The estimate for $f\in C(\bar{D})$ follows by approximation with a sequence $\{f_n;n\in\mathbb{N}\}$ in $C_0(\bar{D})$ such that $|f_n|\leqq|f|$ and measure $\{x\in\bar{D}\,|\,f_n(x)\neq f(x)\}\to 0$ for $n\to\infty$. Since

$$|(L-\lambda)^{-1}f(x)|\leqq\|(L-\lambda)^{-1}f_n\|+\left|\int_D G_\lambda(x,\xi)(f-f_n)(\xi)\,d\xi\right|$$

with G_λ the Green kernel of $(L-\lambda)^{-1}$, it is then a consequence of the integrability of the singularity of the Green kernel.

Step 2. $\exists\sigma_0\geqq 0\exists C>0\exists\varepsilon>0\ \forall\lambda$ with $\operatorname{Re}\lambda\geqq\sigma_0$:

$$|P_i(L-\lambda)^{-1}c_j|\leqq C(1+|\lambda|)^{-\varepsilon}\|c_j\|_Y,\qquad 1\leqq i,j<p.\tag{4.6}$$

In order to prove this estimate, we rewrite $P_i(L-\lambda)^{-1}c_j$ as $P_iA^{-\beta}\cdot A^{\delta}(L-\lambda)^{-1}e_j$ with A as in §2.3, $\delta=\beta-\alpha<1$ and $e_j=A^{\alpha}c_j\in C(\bar{D})$. Using (2.2.9) and (2.2.10), it is not difficult to see that $(L-\lambda)^{-1}e_j$ is an element $C_0D(\gamma)$ with $\delta<\gamma<1$. Now we can estimate

$$\begin{aligned}\left|P_i(L-\lambda)^{-1}c_j\right| &\leqq \left\|P_iA^{-\beta}\right\|'\left\|A^{\delta}(L-\lambda)^{-1}e_j\right\| \\ &\leqq k\left\|P_iA^{-\beta}\right\|'\left\|A^{\gamma}(L-\lambda)^{-1}e_j\right\|^{\delta/\gamma}\left\|(L-\lambda)^{-1}e_j\right\|^{1-\delta/\gamma}\end{aligned}$$

with $\|\cdot\|'=\|\cdot\|_{C_0(\bar{D})'}$.

The second inequality is a consequence of Krasnoselskii (1976, Thm. 14.2, pp. 290).

Next we notice, that $\|A^{\gamma}(L-\lambda)^{-1}e_j\|=\|A^{-\tilde{\gamma}}A(A+\tilde{\lambda})^{-1}\tilde{e}_j\|$ with $\tilde{\gamma}=(1-\gamma)/2$, $\tilde{\lambda}=\lambda-\lambda_0$ and $\tilde{e}_j=A^{-\tilde{\gamma}}e_j$. It is now easy to check that $\|A(A+\tilde{\lambda})^{-1}\tilde{e}_j\|\leqq(K+1)\|\tilde{e}_j\|$ if $\lambda_0\geqq\sigma_0$ with σ_0 as in (4.5).

Consequently $\|A^{\gamma}(L-\lambda)^{-1}e_j\|$ is bounded independently of λ for $\mathrm{Re}\lambda\geqq\sigma_0$.

Hence, (4.5) implies (4.6) with $\varepsilon=1-\delta/\gamma$.

Step 3. A trivial consequence of (4.6) is that $\exists\sigma_0>0$ such that $\forall\lambda$ with $\mathrm{Re}\lambda\geqq\sigma_0$, $\Omega(\lambda)$ as defined in §3, (3.5) is nonsingular.

Next using (3.6) and the results from the previous steps we find that $\forall\lambda$ with $\mathrm{Re}\lambda\geqq\sigma_0$ the resolvent $(L+\Pi-\lambda)^{-1}$ is well defined on Y and satisfies the estimate (4.4).

$$\left\|\left|(L+\Pi-\lambda)^{-1}\right|\right\|_Y\leqq K(1+|\lambda|)^{-1}. \tag{4.7}$$

Using (2.2.9), (2.2.10) it is easily verified that $(L+\Pi-\lambda)^{-1}$ maps $Y_0\subset Y$ into Y_0. Then (4.4) is the restriction of (4.7) to Y_0.

In order to complete the proof that $L+\Pi$ generates an analytic subgroup on Y_0 we shall now use the famous characterisation for generators of analytic semigroups as given in Krasnoselskii (1976, Thm. 13.2, pp. 270). Besides (4.4) this characterisation requires that $L+\Pi-\lambda_0$ is densely defined on Y_0. In order to show this it is sufficient to restrict ourselves to the case $\alpha=0$, for the domain $V(\alpha,\Pi)=\{u\in CD(\alpha+1)\,|\,(L_0-\sigma+\Pi)u\in C_0D(\alpha)\}=A^{-\alpha}V(0,\tilde{\Pi})$ with $\tilde{\Pi}=A^{\alpha}\Pi A^{-\alpha}$. Now take $u_0\in C_0(\bar{D})\cap C^{\infty}(\bar{D})$; then $f_0=(L+\Pi-\lambda_0)u_0$ can be approximated by a smooth function with compact support $\tilde{f}_0$ in the sense of $\|\ \|_{L^q(D)}$ with q arbitrarily large. Now take q so large that because of Sobolev's imbedding theorems $C^{\tilde{\beta}}(\bar{D})$ is compactly imbedded in $W^{2,q}(D)$ with $\tilde{\beta}>\beta$ (see Adams (1975)). Using (3.6) and Agmon, Douglis, Nirenberg's a priori estimates we find that for $\tilde{u}_0=(L+\Pi-\lambda_0)^{-1}\tilde{f}_0$ it holds that $\|u_0-\tilde{u}_0\|_{W^{2,q}(\bar{D})}\leqq C\|f-\tilde{f}_0\|_{L^q(D)}$. The conclusion is that indeed $V(0,\Pi)$ is a dense subset of Y_0.

Using that $(L+\Pi-\lambda)^{-1}$ and $e^{(L+\Pi)t}$, $t>0$ are compact operators on Y the contents of (4.3) are a direct consequence of Hale (1971, Lemmas 22.1 and 22.2). □

The compactness of the resolvent $(L+\Pi-\lambda)^{-1}$ on Y_0 for λ with $\mathrm{Re}\lambda\geqq\sigma_0$ in combination with the compactness of $e^{(L+\Pi)t}$ on Y_0 for $t>0$ and Hale (1971, Lemma 22.1) has another nice implication for the location of the spectrum $\sigma(L+\Pi)$. Namely: if (4.1a) or (4.1b) is satisfied then we are in case 1° mentioned in Theorem 3.1 and moreover $\forall\mu\in\mathbb{R}$

$$\mathbb{X}(\mu)=\{\lambda\in\sigma(L+\Pi)\,|\,\mathrm{Re}\lambda\geqq\mu\}<\infty. \tag{4.8}$$

Equation (4.8) implies that the spectrum $\sigma(L+\Pi)$ can be given as a sequence $\{\lambda_n;n\in\mathbb{N}\}$ such that $n>m\Rightarrow\mathrm{Re}\lambda_n\leqq\mathrm{Re}\lambda_m$ and $\mathrm{Re}\lambda_n\downarrow-\infty$ for $n\uparrow\infty$.

Hence, certainly in this situation the super-singular case cannot occur (see also Kato) (1966, pp. 250(b)).

In §5 more detailed information on the location of $\sigma(L+\Pi)$ will be derived.

5. Some remarks on the location of the spectrum. Again we restrict ourselves to cases where Π falls into one of the classes of (3.3). Let us start with a few simple consequences of the Weinstein–Aronszajn theory described in §3.

LEMMA 5.1. *If* $\lambda_0 \in \sigma(L)$ *satisfies* $m(\lambda_0; L) > q = \dim \operatorname{ran} \Pi$ *then* $\lambda_0 \in \sigma(L+\Pi)$.

Proof of Lemma 5.1. If $\dim \operatorname{ran} \Pi = q$ then $\Pi = \sum_{i=1}^{q} \hat{c}_i \hat{P}_i$ with linearly independent $\hat{c}_i$, $i=1,\cdots,q$ and $\hat{P}_i$, $i=1,\cdots,q$. It is now easy to verify that the last m columns of the matrix $\hat{\Omega}(\lambda_0)$ given in (3.5) are linearly dependent. Hence (3.11)(ii) yields $\lambda_0 \in \sigma(L+\Pi)$. □

LEMMA 5.2. *Suppose* $\lambda \notin \sigma(L)$. *If one of the following conditions holds*

$$\text{(5.1)} \qquad \text{a.} \quad \sum_{i=1}^{p} \left| P_i (L-\lambda)^{-1} c_j \right| < 1 \quad \text{for } j=1,\cdots,p \quad \text{or}$$

$$\text{b.} \quad \sum_{j=1}^{p} \left| P_i (L-\lambda)^{-1} c_j \right| < 1 \quad \text{for } i=1,\cdots,p,$$

then $\lambda \notin \sigma(L+\Pi)$.

Proof. It follows from Gershgorin's theorem (see Wilkinson (1965)) that 0 cannot be an eigenvalues of the matrix $\Omega(\lambda)$ given in (3.6). This implies $\lambda \notin \sigma(L+\Pi)$ because of (3.11)(i). □

The latter lemma appears to be very useful in the derivation of the following theorem.

THEOREM 5.1. *In the case where* Π *lies in the class specified in* (4.1a) *i.e.* $P_i \in HD(\beta)'$, $c_i \in HD(\alpha)$, $i=1,\cdots,p$; $0 \leqq \alpha \leqq \beta < \alpha+1$ *the following estimate holds true for the location of* $\sigma(L+\Pi)$:

$$\text{(5.2)} \qquad \sigma(L+\Pi) \subset \left\{ \lambda \mid d(\lambda) \leqq N |\lambda - \mu_0|^{\gamma} \right\} \cup \left\{ \lambda \mid d(\lambda) \leqq N^{\delta} \right\}$$

with $\gamma = \beta - \alpha$, $\delta = (1-\gamma)^{-1}$, $\mu_0 = \max_{\bar{D}} a_0 + 1$ *and*

$$N = 2p \max_{\substack{1 \leqq i \leqq p \\ 1 \leqq j \leqq p}} \|c_j\|_{HD(\alpha)} \|P_i A^{-\gamma}\|_{HD(\alpha)'},$$

$$d(\lambda) = \text{distance}\,(\lambda, \sigma(L)).$$

Proof of Theorem 5.1. We start with the following observation:

$$\text{(5.3)} \qquad P_i (L-\lambda)^{-1} c_j = \sum_{n \geqq 1} (\mu_n - \lambda)^{-1} (\mu_0 - \mu_n)^{\gamma} P_i A^{-\gamma} \hat{\phi}_n \langle \hat{\phi}_n, c_j \rangle_{HD(\alpha)}$$

with $\hat{\phi}_n = \phi_n / \|\phi_n\|_{HD(\alpha)}$, $n \in \mathbb{N}$.

Since $P_i A^{-\gamma} \in HD(\alpha)'$ there exists an element $W_i \in HD(\alpha)$ with $\|W_i\|_{HD(\alpha)} = \|P_i A^{-\gamma}\|_{HD(\alpha)'}$ such that $P_i A^{-\gamma} = \langle W_i, \cdot \rangle_{HD(\alpha)}$. As a consequence we can deduce from (5.3) the following estimate:

$$\text{(5.4)} \qquad \left| P_i (L-\lambda)^{-1} c_j \right| \leqq \max_{\mu \in \sigma(L)} \frac{|\mu_0 - \mu|^{\gamma}}{|\mu - \lambda|} \sum_{n \geqq 1} \left| \langle W_i, \hat{\phi}_n \rangle_{HD(\alpha)} \right| \left| \langle \hat{\phi}_n, c_j \rangle_{HD(\alpha)} \right|.$$

An application of Schwarz' inequality leads to

$$(5.5)\qquad \sum_{n\geqq 1}\left|\langle W_i,\hat{\phi}_n\rangle_{HD(\alpha)}\right|\left|\langle\hat{\phi}_n,c_j\rangle_{HD(\alpha)}\right|$$

$$\leqq\left\{\sum_{n\geqq 1}\left|\langle W_i,\hat{\phi}_n\rangle_{HD(\alpha)}\right|^2\right\}^{1/2},\qquad\left\{\sum_{n\geqq 1}\left|\langle\hat{\phi}_n,c_j\rangle_{HD(\alpha)}\right|^2\right\}^{1/2}$$

$$=\|W_i\|_{HD(\alpha)}\|c_j\|_{HD(\alpha)}=\|P_iA^{-\gamma}\|_{HD(\alpha)}\|c_j\|_{HD(\alpha)}\leqq\frac{N}{2p}.$$

Further it is clear that for each $\mu\in\sigma(L)$:

$$(5.6)\qquad \frac{|\mu_0-\mu|^\gamma}{|\mu-\lambda|}\leqq\frac{|\mu_0-\lambda|^\gamma+|\mu-\lambda|^\gamma}{|\mu-\lambda|}\leqq\frac{|\mu_0-\lambda|^\gamma}{d(\lambda)}+d(\lambda)^{\gamma-1}.$$

From (5.4)–(5.6) we deduce that for $j=1,\cdots,p$:

$$(5.7)\qquad \sum_{i=1}^{p}\left|P_i(L-\lambda)^{-1}c_j\right|\leqq\frac{N}{2}\left\{\frac{|\mu_0-\lambda|^\gamma}{d(\lambda)}+d(\lambda)^{\gamma-1}\right\}.$$

If $N|\mu_0-\lambda|^\gamma/d(\lambda)<1$ and $Nd(\lambda)^{\gamma-1}<1$ we can conclude from Lemma 5.2 that $\lambda\notin\sigma(L+\Pi)$ i.e. $\sigma(L+\Pi)\subset\{\lambda|N|\mu_0-\lambda|^\gamma/d(\lambda)\geqq 1$ or $Nd(\lambda)^{\gamma-1}\geqq 1\}$ which are exactly the contents of (5.2). □

In some important cases it is possible to improve Theorem 5.1.

THEOREM 5.2. *Suppose* Π *lies in the class specified in* (4.1a) *with* $\beta=\alpha$, (*i.e.* $\gamma=0$ *in* (5.2)). *Then there exists a sequence* $\{\rho_n;n\in\mathbb{N}\}$ *with* $\rho_n\downarrow 0$ *for* $n\uparrow\infty$ *such that*

$$(5.8)\qquad \sigma(L+\Pi)\subset\bigcup_{n=1}^{\infty}\{\lambda\,|\,|\lambda-\mu_n|\leqq\rho_n\}.$$

Explicit expressions for the ρ_n*'s are given in* (5.13).

Proof of Theorem 5.2. Define:

$$(5.9)\qquad N_{ij}(\lambda)=\sum_{\substack{n\in\mathbb{N}\\ \mu_n-\mu_1\leqq\frac{1}{2}(\mathrm{Re}\,\lambda-\mu_1)}}\left|P_i\hat{\phi}_n\right|\left|\langle\hat{\phi}_n,c_j\rangle_{HD(\alpha)}\right|,\qquad N_{ij}=N_{ij}(\mu_1),$$

$$N(\lambda)=2p\max_{\substack{1\leqq i\leqq p\\ 1\leqq j\leqq p}}N_{ij}(\lambda),\qquad N=N(\mu_1).$$

It is not difficult to verify that N as defined here has the same interpretation as the N of Theorem 5.1. It will further be clear that $0\leqq N_{ij}(\lambda)\leqq N_{ij}$, $0\leqq N(\lambda)\leqq N$ and $\mathrm{Re}\,\lambda\downarrow-\infty$ implies $N_{ij}(\lambda)\downarrow 0$, $N(\lambda)\downarrow 0$. Now we have

$$(5.10)\qquad \left|P_i(L-\lambda)^{-1}c_j\right|\leqq\left(\sum_1+\sum_2\right)|\lambda-\mu_n|^{-1}\left|P_i\hat{\phi}_n\right|\left|\langle\hat{\phi}_n,c_j\rangle_{HD(\alpha)}\right|$$

where

Σ_1 runs over $n\in\mathbb{N}$ such that $\mu_n-\mu_1\leqq\frac{1}{2}(\mathrm{Re}\,\lambda-\mu_1)$,

Σ_2 runs over $n\in\mathbb{N}$ such that $\mu_n-\mu_1>\frac{1}{2}(\mathrm{Re}\,\lambda-\mu_1)$.

In Σ_1 we use that $|\lambda-\mu_n|^{-1}\leqq d(\lambda)^{-1}$ and in Σ_2 we use that $|\lambda-\mu_n|^{-1}=|\lambda-\mu_1|^{-1}\cdot|1-(\mu_n-\mu_1)(\lambda-\mu_1)^{-1}|^{-1}\leqq 2|\lambda-\mu_1|^{-1}$. This leads to

$$\left|P_i(L-\lambda)^{-1}c_j\right|\leqq d(\lambda)^{-1}N_{ij}(\lambda)+2|\lambda-\mu_1|^{-1}\left(N_{ij}(\mu_1)\right).$$

So for all $j=1,\cdots,p$

$$\sum_{i=1}^{p}\left|P_i(L-\lambda)^{-1}c_j\right|\leqq\frac{1}{2}d(\lambda)^{-1}N(\lambda)+|\lambda-\mu_1|^{-1}N. \tag{5.11}$$

Using Lemma 5.2 we find from (5.11) that certainly $\lambda\notin\sigma(L+\Pi)$ if

$$d(\lambda)>\frac{3}{2}N(\lambda)\quad\text{and}\quad|\lambda-\mu_1|>\frac{3}{2}N. \tag{5.12}$$

Let us next give a sufficient condition in order to ensure $d(\lambda)>\frac{3}{2}N(\lambda)$. Define $\mu(\lambda)$ as the largest $\mu\in\sigma(L)$ such that $d(\lambda)=|\mu-\lambda|$. It is clear that $\mu(\lambda)=\mu\in\sigma(L)\Leftrightarrow\nu_-(\mu)\leqq\operatorname{Re}\lambda<\nu_+(\mu)$. Here $\nu_+(\mu)=\infty$ if $\mu=\mu_1$ and $\nu_+(\mu)=\frac{1}{2}(\mu+\mu_+)$ else with μ_+ the right-hand neighbour of μ in $\sigma(L)$. Further $\nu_-(\mu)=\frac{1}{2}(\mu+\mu_-)$ with μ_- the left-hand neighbour of μ in $\sigma(L)$. If $\mu(\lambda)=\mu$ then $|\lambda-\mu|>\frac{3}{2}N(\nu_+(\mu))$ is such a sufficient condition.

It is not difficult to verify that (5.12) is certainly satisfied if $\forall n\in\mathbb{N}$

$$|\lambda-\mu_n|>\rho_n=\frac{3}{2}N(\nu_+(\mu_n)), \tag{5.13}$$

and this proves the theorem. □

Some additional remarks to Theorem 5.2 can be made concerning the decay of the ρ_n's for $n\uparrow\infty$.

COROLLARY *to Theorem* 5.2.

a. *If for either* (i): *each c_i consists of a finite linear combination of eigenfunctions ϕ_n or* (ii): *each P_i is of the form $\langle W_i,\cdot\rangle_{L_2(D)}$ where W_i consists of a finite linear combination of eigenfunctions ϕ_n then only finitely many ρ_n's are >0.*

b. *If for $i=1,\cdots,p$: $P_i\in HD(\beta)'$ and $c_i\in HD(\beta+\gamma)$ with $\beta\geqq 0$ and $\gamma>0$ then $\forall n\in\mathbb{N}$:*

$$\rho_n\leqq 3\cdot 2^{\gamma}\cdot p\max_{\substack{1\leqq i\leqq p\\ 1\leqq j\leqq p}}\|P_i\|_{HD(\beta)'}\|A^{\gamma}c_j\|_{HD(\beta)}(\mu_0-\mu_{n+})^{-\gamma} \tag{5.14}$$

i.e. $\rho_n=O(n^{-2\gamma/d})$ for $n\to\infty$.

c. *If for $i=1,\cdots,p$: $P_i=\langle W_i,\cdot\rangle_{HD(\alpha)}$ with $W_i\in HD(\alpha+\gamma)$ and $c_i\in HD(\alpha)$, $\alpha\geqq 0$, $\gamma\geqq 0$ then $\forall n\in\mathbb{N}$.*

$$\rho_n\leqq 3\cdot 2^{\gamma}\cdot p\max_{\substack{1\leqq i\leqq p\\ 1\leqq j\leqq p}}\|A^{\gamma}W_i\|_{HD(\alpha)}\|c_j\|_{HD(\alpha)}(\mu_0-\mu_{n_+})^{-\gamma} \tag{5.15}$$

i.e. $\rho_n=O(n^{-2\gamma/d})$ for $n\to\infty$.

Note that parts b and c can be seen as dual statements.

For part a of the corollary we use (5.13) and the fact that under the given conditions the functions $N_{ij}(\lambda)$ given in (5.9) vanish if $\mathrm{Re}\lambda$ becomes too negative. For part b of the corollary we use (5.13) and the estimate

$$N_{ij}(\lambda) = \sum_{\substack{n\in\mathbb{N} \\ \mu_n-\mu_i\leqq\frac{1}{2}(\mathrm{Re}\lambda-\mu_1)}} (\mu_0-\mu_n)^{-\gamma}|P_i\hat{\phi}_n|\left|\left\langle \hat{\phi}_n, A^{\gamma}c_j\right\rangle_{HD(\beta)}\right|,$$

$$\leqq\left(\frac{1}{2}\right)^{-\gamma}(\mu_0-\mathrm{Re}\lambda)^{-\gamma}\sum_{n\in\mathbb{N}}|P_i\hat{\phi}_n|\left|\left\langle \hat{\phi}_n, A^{\gamma}c_j\right\rangle_{HD(\beta)}\right|.$$

The last statement of the corollary is a consequence of the asymptotics of the eigenvalues. Part c is proven completely analogous to part b.

6. On the occurrence of the super-singular case: $\sigma(L+\Pi)=\mathbb{C}$. First we shall give an example which demonstrates that cases where $\sigma(L+\Pi)=\mathbb{C}$ really can occur.

In this example we take L as in (1.4) and

$$\Pi = c\delta_y L \tag{6.1}$$

with

y some point on ∂D, $\delta_y w = w(y)$,

c some function $\in C^{\infty}(\overline{D})$ with the property,

$c(y) = -1$.

Note that this example falls in the context of §3, for (3.3a) is applicable for any $s > d/2$, $s \notin -\frac{1}{2}+\mathbb{N}$ with d the dimension of the space or (3.3b) is applicable for any $\alpha > 0$, $\alpha \in \mathbb{N}$. Let us now calculate the Weinstein–Aronszajn determinant

$$\omega(\lambda) = 1+\delta_y L(L-\lambda)^{-1}c = 1+\delta_y\left\{1+\lambda(L-\lambda)^{-1}\right\}c \tag{6.2}$$

$$= 1+\delta_y c = 0,$$

for all $\lambda \notin \sigma(L)$! Here we used that $\delta_y(L-\lambda)^{-1}=0$ because of the Dirichlet b.c. and $y \in \partial D$. Consequently $\omega \equiv 0$ on $\mathbb{C}$ and Theorem 5.1 yields $\sigma(L+\Pi)=\mathbb{C}$.

We observe that in this example the operator L is not densely defined on the space Y. Hence it is not very surprising, that perturbations of L can destroy its nice spectral properties. Nevertheless, even in this situation though apparently cases with $\sigma(L+\Pi)=\mathbb{C}$ can occur, one intuitively gets the feeling that the occurrence of these super-singular cases is very rare. This feeling will be given a mathematical base in Theorem 7.1. First we introduce some concepts. From now on we suppose that X, Y are spaces of real or complex functions as indicated in (3.3).

The set of admissible feedback controls is defined to be

$$RC = \{\Pi \in \mathscr{L}(X \to Y) \mid \Pi \text{ real and } \dim \mathrm{ran}\,\Pi \leqq p\}. \tag{6.3}$$

RC is a closed subset of $\mathscr{L}(X \to Y)$, which we endow with the norm inherited from $\mathscr{L}(X \to Y)$. Note that RC is not a linear space. It is true that $\Pi \in RC \Rightarrow \alpha\Pi \in RC$ for each $\alpha \in \mathbb{R}$, but $\Pi_1 \in RC$ and $\Pi_2 \in RC$ does not necessarily imply $\Pi_1+\Pi_2 \in RC$!

Each element of RC is of the form $\sum_{i=1}^{p} c_i P_i$ with some real $(c,P) \in Y^p \times (X')^p$. However quite a number of different (c,P)'s in $Y^p \times (X')^p$ correspond to the same $\Pi \in RC$. This induces an equivalence relation on $Y^p \times (X')^p$ as follows $(c,P) \sim (\hat{c},\hat{P}) \Leftrightarrow \sum_{i=1}^{p} c_i P_i = \sum_{i=1}^{p} \hat{c}_i \hat{P}_i$.

So RC can be identified with $\{Y^p \times (X')^p\}/\sim$. The following statement, which relates the topology of RC to the topology of $Y^p \times (X')^p$, can easily be verified. Let $\Pi \in RC$ be given with $\dim \operatorname{ran} \Pi = q \leqq p$, $\Pi = \sum_{i=1}^{p} c_i P_i$ with $c_i = 0$, $P_i = 0$ for $i > q$. Then $\exists M > 0\ \exists \varepsilon_0 > 0\ \forall \varepsilon \in (0, \varepsilon_0)\ \forall \Pi^{(1)} \in RG$:

$$\|\Pi - \Pi^{(1)}\|_{RC} \leqq \varepsilon \Rightarrow \exists (c^{(1)}, P^{(1)}) \in Y^p \times (X')^p, \tag{6.4}$$

such that $\|(c, P) - (c^{(1)}, P^{(1)})\|_{Y^p \times (X')^p} \leqq M\varepsilon^{1/2}$ and $\Pi^{(1)} = \sum_{i=1}^{p} c_i^{(1)} P_i^{(1)}$.

Let us further introduce RC_0, RC_s the subsets of RC consisting of nonsingular, singular controls

$$\begin{aligned} &\Pi \in RC_0 \Rightarrow \sigma(L + \Pi) \neq \mathbb{C}, \\ &\Pi \in RC_s \Rightarrow \sigma(L + \Pi) = \mathbb{C}. \end{aligned} \tag{6.5}$$

THEOREM 6.1.

a. *The property of nonsingularity of the feedback control is generic*: RC_0 *is an open and dense subset of* RC.

b. *Singular feedback controls have a neglectable probability of occurrence in the following sense*: *for each finite dimensional linear subspace* H *of* RC *the Lebesgue measure in* H *of* $H \cap RC_s$ *equals* 0.

Proof of Theorem 6.1. In order to express the dependence of the Weinstein–Aronszajn determinant on Π we introduce the notation $\omega(\lambda; \Pi)$. Using (6.4) it is easily checked that for each fixed $\lambda \notin \sigma(L)$ the map $\omega(\lambda; \cdot)$ from RC into $\mathbb{C}$ is continuous.

(i) We shall show that RC_0 is open in RC. $\Pi \in RC_0$ implies $\omega(\lambda; \Pi) \neq 0$ for some $\lambda \notin \sigma(L)$, see Theorem 5.1. Consequently the following formula holds true:

$$RC_0 = \bigcup_{\lambda \notin \sigma(L)} \omega(\lambda; \cdot)^{-1}\{\mathbb{C} \setminus \{0\}\}.$$

$\mathbb{C} \setminus \{0\}$ is open and $\omega(\lambda; \cdot)^{-1}\{\mathbb{C} \setminus \{0\}\}$ is open in RC. RC_0 is the union of open sets and hence it is open itself.

(ii) Let us next demonstrate that RC_0 is dense in RC. Take $\Pi \in RC_0$. For fixed $\lambda \notin \sigma(L)$ we have that $\omega(\lambda; (1 - \varepsilon)\Pi)$ is a nonconstant polynomial in ε. So for $\varepsilon \neq 0$ and ε arbitrarily small we see that $(1 - \varepsilon)\Pi \in RC_0$.

(iii) Our last step is to show that for each finite dimensional linear subspace H of RC we have meas $(H \cap RC_s) = 0$. Let $\Pi_1, \cdots, \Pi_m$ be a base of H. For each fixed $\lambda \notin \sigma(L)$ the function $\omega(\lambda; \sum_{i=1}^{m} \xi_i \Pi_i)$ defines a polynomial in $\xi_1, \cdots, \xi_m$ which equals 1 for $\xi = 0 \in \mathbb{R}^m$. This implies that the set $\{\xi \in \mathbb{R}^m \mid \omega(\lambda; \sum_{i=1}^{m} \xi_i \Pi_i) = 0\}$ has measure $= 0$ in $\mathbb{R}^m$. In other words $Z(\lambda; H) = \{\Pi \in H \mid \omega(\lambda; \Pi) = 0\}$ has measure $= 0$ in H. So certainly $RC_s \cap H = \bigcap_{\lambda \notin \sigma(L)} Z(\lambda; H)$ has measure zero in H. □

7. On the completeness of generalized eigenspaces and their eigenprojections. Let us assume that Π is of one of the types indicated in (3.3) and that $\sigma(L + \Pi) \neq \mathbb{C}$. The resolvent $(L + \Pi - \lambda)^{-1}$ is then a compact operator on Y for $\lambda \notin \sigma(L + \Pi)$.

Hence to each $\lambda \in \sigma(L + \Pi)$ there corresponds a generalized eigenspace $N(\lambda)$ with a dimension equal to the algebraic multiplicity of λ and a projection operator $P(\lambda)$ into $N(\lambda)$. Here $N(\lambda)$ and $P(\lambda)$ with $\lambda \in \sigma(L + \Pi)$ are given by:

$$N(\lambda) = \text{null space of } (L + \Pi - \lambda)^s, \tag{7.1}$$

$$P(\lambda) = \frac{1}{2\pi i} \int_{\Gamma(\lambda)} (L + \Pi - \lambda)^{-1} d\zeta, \tag{7.2}$$

see Kato (1966, Chap. III).

In (7.1) s is a sufficiently large number $\in \mathbb{N}$. The smallest number $s \in \mathbb{N}$ such that (7.1) is true will be denoted by $s(\lambda)$. In (7.2) $\Gamma(\lambda)$ is a smooth contour which encircles the eigenvalue λ in such a way that $\sigma(L+\Pi)\backslash\{\lambda\}$ lies outside $\Gamma(\lambda)$.

In Lemma 7.1 we shall give somewhat more explicit expressions for the forms of $N(\lambda)$ and $P(\lambda)$, $\lambda \in \sigma(L+\Pi)$.

Further, in Theorems 7.1 and 7.2 we shall prove that under certain conditions the system of eigenprojections is complete i.e.

$$I = \sum_{\lambda \in \sigma(L+\Pi)} P(\lambda).$$

Let us now first introduce some further notation. For $\lambda \in \sigma(L+\Pi)$ we define $q(\lambda) \in \mathbb{Z}$ and $p \times p$ matrices $Q_r(\lambda)$ with $r \in \mathbb{Z}$, $r \geqq q(\lambda)$ by

$$\Omega(\zeta)^{-1} = \sum_{r \geqq q(\lambda)} (\zeta - \lambda)^r Q_r(\lambda). \tag{7.3}$$

Here (7.3) expresses the Laurent expansion for $\zeta \to \lambda$ of the matrix function $\Omega(\zeta)^{-1}$ with Ω as defined in (3.5). $q(\lambda)$ is such that $Q_{q(\lambda)}(\lambda) \neq 0$, so $q(\lambda)$ is the Laurent index of Ω^{-1} at the point λ.

From the Weinstein–Aronszajn theory it follows, that if $\lambda \in \sigma(L+\Pi)\backslash\sigma(L)$ then $q(\lambda) < 0$.

In Lemma 7.1 the notation $\langle\ ,\ \rangle$ will be used analogous to (3.6) and (3.9).

LEMMA 7.1.

a. *If* $\lambda \in \sigma(L+\Pi)\backslash\sigma(L)$ *then*

(i) $N(\lambda)$ *has a basis* $\{\chi_i^\lambda : 1 \leqq i \leqq \dim N(\lambda)\}$ *with each* χ_i^λ *a linear combination of the functions* $(L-\lambda)^{-k} c_j$, $1 \leqq k \leqq s(\lambda)$, $1 \leqq j \leqq p$;

(ii) *the eigenprojection* $P(\lambda)$ *has the form*

$$P(\lambda) = \sum_{\substack{r \geqq q(\lambda) \\ k \geqq 1,\ m \geqq 1 \\ k+m+r=1}} \left\langle (L-\lambda)^{-k} c, Q_r(\lambda) P (L-\lambda)^{-m} . \right\rangle. \tag{7.4}$$

b. *If* $\lambda \in \sigma(L+\Pi) \cap \sigma(L)$ *then*

(i) $N(\lambda)$ *has a basis* $\{\chi_i^\lambda; 1 \leqq i \leqq \dim N(\lambda)\}$ *with each* χ_i^λ *a linear combination of the functions* $(L-\lambda)_*^{-k} c_j$, ϕ_n, $1 \leqq k \leqq s(\lambda)$, $1 \leqq j \leqq p$, $n \in J(\lambda)$;

(ii) *the eigenprojection* $P(\lambda)$ *has the form*

$$\begin{aligned} P(\lambda) = E(\lambda) &- \sum_{\substack{r \geqq q(\lambda) \\ k \geqq 1 \\ k+r=1}} \left\langle (L-\lambda)_*^{-k} c, Q_r(\lambda) P E(\lambda) . \right\rangle \\ &- \sum_{\substack{r \geqq q(\lambda) \\ m \geqq 1 \\ m+r=1}} \left\langle E(\lambda) c, Q_r(\lambda) P (L-\lambda)_*^{-m} . \right\rangle \\ &+ \sum_{\substack{r \geqq q(\lambda) \\ k \geqq 1,\ m \geqq 1 \\ k+m+r=1}} \left\langle (L-\lambda)_*^{-k} c, Q_r(\lambda) P (L-\lambda)_*^{-m} . \right\rangle. \end{aligned} \tag{7.5}$$

Proof of Lemma 7.1.

(i) Let us first prove a(i) and b(i). Let us denote by $(L+\Pi-\lambda)^{-1} g$ with $\lambda \in \sigma(L+\Pi)$ the space of solutions u of (3.1)–(3.2). In case a $(L+\Pi-\lambda)^{-1} g$ is given by the right-hand side of (3.6) with $\Omega(\lambda)^{-1} P (L-\lambda)^{-1} g$ interpreted as the set of solutions of

$\Omega(\lambda)\xi = P(L-\lambda)^{-1}g$. In case b $(L+\Pi-\lambda)^{-1}g$ is given by the right-hand side of (3.9) where $\hat{\Omega}(\lambda)^{-1}\binom{P(L-\lambda)^{-1}_*}{\phi'}g$ is interpreted in the obvious way. Of course for $V \subset Y$ we mean by $(L+\Pi-\lambda)^{-1}V$ the union of the $(L+\Pi-\lambda)^{-1}$ with $g \in V$.

Now we have the formula

$$N(\lambda) = (L+\Pi-\lambda)^{-s(\lambda)}\{0\}. \tag{7.6}$$

Repeated application of the adapted forms of (3.6)–(3.9) to (7.6) shows immediately that a(i) and b(i) hold.

(ii) Let us now prove a(ii) and b(ii). From the definition in (7.2) we derive using the expression (3.6) for $(L+\Pi-\zeta)^{-1}$

(7.7)

$$P(\lambda) = -\frac{1}{2\pi i}\oint_{\Gamma(\lambda)}(L-\zeta)^{-1}\,d\zeta + \frac{1}{2\pi i}\oint_{\Gamma(\lambda)}\left\langle (L-\zeta)^{-1}c, \Omega(\zeta)^{-1}(L-\zeta)^{-1}\cdot\right\rangle d\zeta.$$

For $\Gamma(\lambda)$ we choose a contour $|\lambda-\zeta| = \delta$ with $\delta > 0$ so small that $\{\sigma(L+\Pi)\cup\sigma(L)\}\cap\{\zeta \,|\, |\lambda-\zeta| \leqq \delta\} = \{\lambda\}$.

Next we plug the Laurent expansion of $\Omega(\zeta)^{-1}$ given in (8.3) and the Laurent expansion of $(L-\zeta)^{-1}$ for $\zeta \to \lambda$ to be given here below in (7.8) into (7.7):

$$\begin{aligned}(L-\zeta)^{-1} &= \sum_{k=0}^{\infty}(L-\lambda)^{-k-1}(\zeta-\lambda)^k \quad \text{in case a,}\\ -(\zeta-\lambda)^{-1}E(\lambda) &+ \sum_{k=0}^{\infty}(L-\lambda)_*^{-k-1}(\zeta-\lambda)^k \quad \text{in case b.}\end{aligned} \tag{7.8}$$

The convergence of the sums in (7.8) is in $L(HD(\alpha) \to HD(\alpha))$. An application of the Cauchy residue theorem then gives (7.4) in case a and (7.5) in case b. □

Let us now suppose that Π satisfies the same assumption as in (4.1a), namely for $i = 1, \cdots, p$: $c_i \in HD(\alpha)$, $P_i \in HD(\beta)$, $0 \leqq \alpha \leqq \beta < \alpha + 1$.

The spectrum $\sigma(L+\Pi)$ can then be given as a sequence $\{\lambda_n; n \in \mathbb{N}\}$ such that $n > m \Rightarrow \lambda_m$, $\mathrm{Re}\lambda_n \leqq \mathrm{Re}\lambda_m$ and $\mathrm{Re}\lambda_n \downarrow -\infty$ for $n \uparrow \infty$.

Let us denote for a finite subset $F \subset \mathbb{N}$ by $N(\lambda_n; n \in F)$ the finite-dimensional linear subspace of $HD(\alpha)$ spanned by the $N(\lambda_n)$'s with $n \in F$.

Then since A is selfadjoint, $A^{-1}\Pi$ is compact and A^{-m} is of Hilbert type, it is a well-known result (see Dunford and Schwartz (1962, p. 2374) or Gohberg and Krein (1969, pp. 276–277)) that the generalized eigenspaces are complete in the following sense

$$\begin{aligned}&\forall u \in HD(\alpha)\ \exists\{u_k; k \in \mathbb{N}\} \text{ with } u_k \in N(\lambda_n; 1 \leqq n \leqq k)\\ &\text{such that } \lim_{k\to\infty}\|u-u_k\|_{HD(\alpha)} = 0.\end{aligned} \tag{7.9}$$

This result shows that each element of $HD(\alpha)$ can be approximated with linear combinations of generalized eigenfunctions.

Next we shall prove that the eigenprojections $P(\lambda)$ with $\lambda \in \sigma(L+\Pi)$ are "conditionally complete" in certain cases.

This means that elements of $HD(\alpha)$ can be expanded in terms of generalized eigenfunctions. In order to prove this, we have to make rather strong assumptions. In the first place:

$$\text{for } 1 \leqq i \leqq p\text{: } P_i \in HD(\alpha)', \qquad c_i \in HD(\alpha) \quad \text{with } \alpha \geqq 0, \tag{7.10}$$

i.e. Π falls in the class of (4.1a) with $\alpha=\beta$. For our other assumptions we have to introduce some notation. Let $\{\hat{\phi}_n; n\in\mathbb{N}\}$ be the sequence of eigenfunctions of the uncontrolled problem normalized in $HD(\alpha)$ i.e. $\hat{\phi}_n=\phi_n/\|\phi_n\|_{HD(\alpha)}$, ϕ_n as in (2.1.1), (2.1.2).

For $n\in\mathbb{N}$ we define the following numbers:

$$C^{(n+1)}=\sqrt{p}\max_{j=i,\cdots,p}\left\{\sum_{\substack{k\in\mathbb{N}\\ \mu_k-\mu_1\leqq\frac{1}{2}(\mu_n-\mu_1)}}\left|\left\langle c_j,\hat{\phi}_k\right\rangle_{HD(\alpha)}\right|^2\right\}^{1/2},$$

$$Q^{(n+1)}=p\max_{i=1,\cdots,p}\left\{\sum_{\substack{k\in\mathbb{N}\\ \mu_k-\mu_1\leqq\frac{1}{2}(\mu_n-\mu_1)}}\left|P_i\hat{\phi}_k\right|^2\right\}^{1/2}, \tag{7.11}$$

$$r(n)=4C^{(n)}Q^{(n)},$$

$$h(n)=\frac{r(n-1)}{\mu_{n-1}-\mu_n},\qquad g(n)=\frac{(\mu_n-\mu_{n+1})(\mu_1-\mu_n)}{\left(C^{(n)}+Q^{(n)}\right)^2}.$$

Here we use the convention: $a.0^{-1}=\infty$ for $a>0$ and $0.0^{-1}=1$. Note that $r(n)\downarrow 0$ for $n\uparrow\infty$ and, that $r(n)\geqq\rho_n$ with ρ_n as in Theorem 5.2 and (5.13), if $\mu_n<\mu_{n-1}$. Consequently we have $\sigma(L+\Pi)\subset\bigcup_{n\in\mathbb{N}}\{\lambda\,|\,|\mu_n-\lambda|\leqq r(n)\}$. This fact will play an important role further on. In order to show the conditional completeness of the eigenprojections, we now assume that there exists a strictly increasing subsequence of $\mathbb{N}$, which we denote by $\{s(n); n\in\mathbb{N}\}$, with the following properties:

$$\lim_{n\to\infty}h(s(n))=0, \tag{7.12}$$

$$\lim_{n\to\infty}g(s(n)-1)=\infty. \tag{7.13}$$

These conditions are rather frequently satisfied. Roughly speaking they require that the expansion coefficients of the c_i's and P_j's w.r.t. the basis of eigenfunctions corresponding to L decay sufficiently fast compared with the length of certain gaps in the spectrum of L towards $+\infty$. For example, in the case of the Laplace operator: $L=\Delta$ on a block: $D=\{x\,|\,|x_i-a_i|<L_i, i=1,\cdots,d\}$ in d dimensions, these assumptions are fulfilled, because the difference between two different consecutive eigenvalues $\mu_{n-1}-\mu_n$ is larger than π^2L^{-2} with $L=\min_{i=1,\cdots,d}L_i$, and: $\mu_1-\mu_n\uparrow\infty$ for $n\uparrow\infty$, $C^{(n)}$, $Q^{(n)}\downarrow 0$ for $n\uparrow\infty$. From the general theory for the asymptotics of the eigenvalues given in (2.1.5) it follows that:

$$\exists K>0\ \forall M>0\ \exists n\geqq M \text{ such that}$$
$$\text{(i) } \mu_1-\mu_n\leqq Kn^{2/d} \text{ and (ii) } \mu_{n-1}-\mu_n\geqq Kn^{(2-d)/d}$$

with d the dimension of the domain D. The second part of this statement has to hold, because its denial: $\forall K>0\ \exists M>0$ such that $\forall n\geqq M$: $\mu_{n-1}-\mu_n\leqq Kn^{(2-d)/d}$, implies: $\mu_1-\mu_n=o(n^{2/d})$ for $n\to\infty$, which contradicts (2.1.5). For a suitable choice of K the first part holds for every n. An immediate consequence of this statement is that

$$\exists K>0\ \forall M>0\ \exists n\geqq M:$$
$$\text{(i) } h(n)\leqq K^{-1}r(n-1)\cdot n^{(d-2)/d} \text{ and (ii) } g(n-1)\geqq K^2n^{(4-d)/d}\cdot R\left(C^{(n)}+Q^{(n)}\right)^{-2}.$$

The conclusion is that (7.12) is automatically satisfied if $d=1,2$ and that (7.13) is automatically satisfied if $d=1,2,3,4$. In order to ensure that (7.12)–(7.13) are satisfied in situations where the dimension of the domain is not as specified above, it is sufficient to require more regularity for the c_i's and P_j's:

$$c_i \in HD(\alpha+\gamma),$$
$$P_i = \langle W_i, \cdot \rangle_{HD(\alpha)} \text{ with } W_i \in HD(\alpha+\gamma).$$

It will be clear that both $C^{(n)}$ and $Q^{(n)}$ behave then as $o((\mu_1-\mu_n)^{-\gamma})$ for $n\to\infty$. Therefore, (7.12) is fulfilled if $\gamma \geqq \max(0, \frac{1}{4}(d-2))$ and in order to fulfill (7.13) we must have $\gamma \geqq \max(0, \frac{1}{4}(d-4))$. Thus we have demonstrated that (7.11)–(7.13) are not very restrictive.

Finally we introduce the notation for $S \subset \mathbf{C}$

$$P(\lambda;\lambda\in S) = \sum_{\lambda\in S\cap\sigma(L+\Pi)} P(\lambda), \tag{7.14}$$
$$P(\lambda;\lambda\in\varnothing)=0.$$

We shall prove the following result.

THEOREM 7.1. *Suppose that the conditions* (7.10) *and* (7.13) *are satisfied. Then the eigenprojections are conditionally complete in the following sense*:

$$I = \sum_{n=1}^{\infty} P(\lambda;\lambda\in S_n) \tag{7.15}$$

where by definition: $S_1=\{\lambda \mid \mathrm{Re}\lambda > \mu_{s(n-1)}+r(1)\}$ *and for* $n>1$ $S_n=\{\lambda \mid \mu_{s(n)}+r(n)< \mathrm{Re}\lambda < \mu_{s(n-1)}+r(n-1)\}$. *The convergence of the sum in* (7.15) *is strongly in* $HD(\alpha)$.

Proof of Theorem 7.1. The scheme of this proof is identical to the proof of Kato (1966, Thm. 4.5, Chap. V, §4.5, pp. 293–295). We already know that

$$I = \sum_{n=1}^{\infty} E(\lambda;\lambda\in S_n) \tag{7.16}$$

where $E(\lambda;\lambda\in S_n)=\sum_{\lambda\in S_n\cap\sigma(L)}E(\lambda)$ with $E(\lambda)$ as in (2.1.4). The convergence of the sum in (7.16) is strongly in $HD(\alpha)$. We shall now show that

$$R_N = \sum_{n=1}^{N} P(\lambda;\lambda\in S_n) - \sum_{n=1}^{N} E(\lambda;\lambda\in S_n) \to 0 \quad \text{for } N\to\infty \tag{7.17}$$

in $\mathscr{L}(HD(\alpha)\to HD(\alpha))$. It is then clear that (7.15) holds true. We have the following formula

$$R_N = -\frac{1}{2\pi i}\oint_{\gamma_N} \{(L+\Pi-\xi)^{-1}-(L-\xi)^{-1}\}\,d\zeta, \tag{7.18}$$

see, Kato (1966, p. 294, 4.19). Here γ_N is the contour consisting of

(i) half the circle $\{\zeta \mid |\xi-\nu_N|=R, \mathrm{Re}\,\xi \geqq \nu_N\}$,

(ii) the vertical interval $\{\xi \mid \mathrm{Re}\,\xi=\nu_N, |\mathrm{Im}\,\xi| \geqq R\}$, with $\nu_N=\frac{1}{2}(\mu_{s(N)}+\mu_{s(N)-1})$ and R sufficiently large. Of course, (7.18) holds since $\{\sigma(L+\Pi)\cup\sigma(L)\}\cap\bigcup_{n=1}^{N}S_n$ lies inside the contour γ_N, if R is sufficiently large (Theorem 5.2!)

Using the form of $(L+\Pi+\xi)^{-1}$ given in (3.6), we get

$$R_N=\frac{1}{2\pi i}\oint_{\gamma_n}\left\langle\left\{(L-\xi)^{-1}c,\Omega(\xi)^{-1}P(L-\xi)^{-1}\right\}\right\rangle d\xi. \tag{7.19}$$

Since on the half circle $\{\xi\,|\,|\xi-\nu_N|=R,\operatorname{Re}\xi\geqq\nu_N\}$ $|||(L-\xi)^{-1}|||_{HD(\alpha)}=O(1/R)$ for $R\uparrow\infty$, we can reduce (7.19) to

$$R_N=\frac{1}{2\pi i}\int_{-\infty}^{\infty}\left\langle(L-\nu_N-i\xi)^{-1}c,\Omega(\nu_N+i\xi)^{-1}P(L-\nu_N-i\xi)^{-1}\right\rangle d\xi. \tag{7.20}$$

Thus we obtain the following estimate

$$|||R_n|||_{HD(\alpha)}\leqq\frac{w_0}{\pi}\int_0^{\infty}\left\|(L-\nu_N-\xi)^{-1}c\right\|_{\{HD(\alpha)\}^p}\left\|P(L-\nu_N-i\xi)^{-1}\right\|_{\{HD(\alpha)'\}^p}d\xi \tag{7.21}$$

with

$$w_0=\max_{\xi\notin\bigcup_{n\in\mathbb{N}}\{\lambda\,|\,|\mu_n-\lambda|\leqq r(n)\}}\left|\left|\left|\Omega(\zeta)^{-1}\right|\right|\right|_{\mathbb{C}^p}.$$

Let us first show that $w_0<\infty$. Analogous to the proof of Theorem 5.2 one can show that $\zeta\in\bigcup_{n\in\mathbb{N}}\{\lambda\,|\,|\mu_n-\lambda|\leqq r(n)\}$ implies for $j=1,\cdots,p$:

$$\begin{aligned}\sum_{i=1}^{p}\left|P_i(L-\zeta)^{-1}c_j\right|&\leqq\frac{1}{d(\zeta)}\sum_{i=1}^{p}\sum_{\substack{n\\ \mu_n-\nu_1\leqq\frac{1}{2}(\operatorname{Re}\zeta-\mu_1)}}|P_i\hat{\phi}_n|\cdot\left|\left\langle\hat{\phi}_n,c_j\right\rangle_{HD(\alpha)}\right|\\&\quad+\frac{2}{|\xi-\mu_1|}\cdot\sum_{i=1}^{p}\sum_{\substack{n\in\mathbb{N}\\ \mu_n-\mu_1>\frac{1}{2}(\operatorname{Re}\xi-\mu_1)}}|P_i\cdot\hat{\phi}_n|\cdot\left|\left\langle\hat{\phi}_n,c_j\right\rangle_{HD(\alpha)}\right|\\&\leqq r(n)^{-1}\cdot\frac{1}{4}r(n)+2\cdot r(1)^{-1}\cdot\frac{1}{4}r(1)=\frac{3}{4}.\end{aligned}$$

So in the matrix norm $\|Z\|_a=\max_{j=1,\cdots,p}\sum_{i=1}^{p}|Z_{ij}|$ it is true that $\|I-\Omega(\zeta)\|_a\leqq\frac{3}{4}$. Consequently $\Omega(\zeta)$ is invertible and $\|\Omega(\zeta)^{-1}\|_a\leqq 4$. Since there is a constant $M>0$, such that $|||\cdot|||_{C^p}\leqq M\|\cdot\|_a$ it is clear that $w_0<\infty$.

In order to get further estimates from (7.21), we use the following trick. We decompose

$$c=E_1^Nc+E_2^Nc,\qquad P=PE_1^N+PE_2^N \tag{7.22}$$

with

$$\begin{aligned}E_1^N&=\sum_{\substack{k\in\mathbb{N}\\ \mu_k-\mu_1\leqq\frac{1}{2}d(N)}}\hat{\phi}_k\left\langle\hat{\phi}_k,\cdot\right\rangle_{HD(\alpha)},\\E_2^N&=\sum_{\substack{k\in\mathbb{N}\\ \mu_k-\mu_1>\frac{1}{2}d(N)}}\hat{\phi}_k\left\langle\hat{\phi}_k,\cdot\right\rangle_{HD(\alpha)},\\d(N)&=\mu_{s(N)-1}-\mu_1.\end{aligned}$$

It is easy to verify that for $\zeta \in (0, \infty)$:

$$
\begin{aligned}
&\left\|(L-\nu_N-i\zeta)^{-1}E_1^N c\right\|_{HD(\alpha)^p} \leqq \left\{\frac{1}{4}i_0(N)^2+\zeta^2\right\}^{-1/2} C_1^N,\\
&\left\|(L-\nu_N-i\zeta)^{-1}E_2^N c\right\|_{HD(\alpha)^p} \leqq \left\{\frac{1}{4}d(N)^2+\zeta^2\right\}^{-1/2} C_2,\\
&\left\|PE_1^N(L-\nu_N-i\zeta)^{-1}\right\|_{\{HD(\alpha)'\}^p} \leqq \left\{\frac{1}{4}i_0(N)^2+\zeta^2\right\}^{-1/2} Q_1^N,\\
&\left\|PE_2^N(L-\nu_N-i\zeta)^{-1}\right\|_{\{HD(\alpha)'\}^p} \leqq \left\{\frac{1}{4}d(N)^2+\zeta^2\right\}^{-1/2} Q_2,
\end{aligned}
\tag{7.23}
$$

with $i_0(N)=\mu_{s(N)-1}-\mu_{s(N)}$.

$$
\begin{aligned}
C_1^N &= \sqrt{p}\max_{j=1,\cdots,p}\left\{\sum_{\substack{k\in\mathbb{N}\\ \mu_k-\mu_1\leqq\frac{1}{2}d(N)}}\left|\left\langle\phi_k, c_j\right\rangle_{HD(\alpha)}\right|^2\right\}^{1/2} = C^{(s(N)-1)},\\
C_2 &= \sqrt{p}\max_{j=1,\cdots,p}\|c_j\|_{HD(\alpha)},\\
Q_1^N &= \sqrt{p}\max_{i=1,\cdots,p}\left\{\sum_{\substack{k\in\mathbb{N}\\ \mu_k-\mu_1\leqq\frac{1}{2}d(N)}}\left|P_i\hat{\phi}_k\right|^2\right\}^{1/2} = Q^{(s(N)-1)},\\
Q_2 &= \sqrt{p}\max_{i=1,\cdots,p}\|P_i\|_{HD(\alpha)'}.
\end{aligned}
$$

Using (7.22)–(7.23), we obtain from (7.21)

(7.24)

$$
\begin{aligned}
\|\!|R_N|\!\|_{HD(\alpha)} &\leqq \frac{w_0}{\pi}\left\{C_1^N Q_1^N\int_0^\infty\left[\frac{1}{4}i_0(N)^2+\zeta^2\right]^{-1}d\zeta\right.\\
&\quad +\left(C_1^N Q_2+C_2 Q_1^N\right)\int_0^\infty\left[\frac{1}{4}i_0(N)^2+\zeta^2\right]^{-1/2}\left[\frac{1}{4}d(N)^2+\zeta^2\right]^{-1/2}d\zeta\\
&\quad \left.+C_2 Q_2\int_0^\infty\left[\frac{1}{4}d(N)^2+\zeta^2\right]^{-1}d\zeta\right\}\\
&\leqq w_0\left\{\frac{C_1^N Q_1^N}{i_0(N)}+\frac{C_1^N Q_2+C_2 Q_1^N}{\sqrt{i_0(N)d(N)}}+\frac{C_2 Q_2}{d(N)}\right\}\\
&\leqq w_0\left\{\frac{1}{4}h(s(N))+g(s(N)-1)^{-1/2}(C_2+Q_2)+\frac{C_2 Q_2}{d(N)}\right\}.
\end{aligned}
$$

Using (7.13) it is clear that $\lim_{N\to\infty}\|\!|R_N|\!\|_{HD(\alpha)}=0$ and this proves the theorem. □

In the following theorem it is shown that at least in certain cases the eigenprojections are not only conditionally complete (i.e. complete for a prescribed way of summation) but even unconditionally complete.

THEOREM 7.2. *Suppose that*:

(i) *only a finite number of eigenvalues* $\lambda \in \sigma(L)$ *have multiplicities* >1;

(ii) *the feedback operator* Π *satifies* (7.10) *and instead of* (7.13) *the following stronger condition holds*:

$$\sum_{n=1}^{\infty} \left| \frac{C^{(n)}}{i(n)} \right|^2 < \infty, \quad \sum_{n=1}^{\infty} \left| \frac{Q^{(n)}}{i(n)} \right|^2 < \infty, \quad \lim_{n \to \infty} \left| \frac{r(n)}{i(n)} \right| = 0 \tag{7.25}$$

with $C^{(n)}$, $Q^{(n)}$, $r(n)$ *as in* (7.11) *and* $i(n)$ = *the distance of* μ_n *to* $\sigma(L) \setminus \{\mu_n\}$;

(iii) *the dimension of the domain D satisfies* $d \leqq 3$.

Then the eigenprojections are unconditionally complete, i.e.

$$I = \sum_{\lambda \in \sigma(L+\Pi)} P(\lambda) \tag{7.26}$$

without any restriction on the numbering in the summation.

Proof of Theorem 7.2. We shall show that there is a constant $M>0$ such that for each subset $\Lambda \subset \sigma(L+\Pi)$

$$\left\| \sum_{\lambda \in \Lambda} P(\lambda) \right\|_{HD(\alpha)} \leqq M. \tag{7.27}$$

Let us first demonstrate that (7.27) implies (7.26). Let $\{\Lambda_n : n \in \mathbb{N}\}$ be a sequence of subsets of $\sigma(L+\Pi)$ such that each Λ_n is finite, $m>n \Rightarrow \Lambda_m \supset \Lambda_n$ and $\sigma(L+\Pi) = \bigcup_{n \in \mathbb{N}} \Lambda_n$.

Take $\varepsilon>0$ arbitrarily small and pick $u \in HD(\alpha)$. Using Theorem 7.1, we can choose $N(\varepsilon)>0$ such that

$$\left\| \left\{ I - \sum_{n=1}^{N(\varepsilon)} P(\lambda; \lambda \in S_n) \right\} u \right\|_{HD(\alpha)} \leqq \varepsilon.$$

Take $M_0(\varepsilon)$ such that $\Lambda_{M_0(\varepsilon)} \supset \bigcup_{n=1}^{N(\varepsilon)} S_n \cap \sigma(L+\Pi)$. Then we have for $m \geqq M_0(\varepsilon)$:

$$I - \sum_{\lambda \in \Lambda_m} P(\lambda) = \left(\sum_{\lambda \in \sigma(L+\Pi) \setminus \Lambda_m} P(\lambda) \right) \left(I - \sum_{n=1}^{N(\varepsilon)} P(\lambda; \lambda \in S_n) \right).$$

So $\| \{ I - \Sigma_{\lambda \in \Lambda_m} P(\lambda) \} u \|_{HD(\alpha)} \leqq M\varepsilon$, i.e. $I = \lim_{n \to \infty} \Sigma_{\lambda \in \Lambda_n} P(\lambda)$ strongly in $HD(\alpha)$. In other words (7.27) implies (7.26).

In order to prove (7.27), our reasoning proceeds as follows. Define $\sigma_1(L)$ as the subset of $\sigma(L)$ consisting of all eigenvalues μ_n which have a multiplicity 1 and for which $r(n)/i(n) < \frac{1}{2}$, $h(n) < \frac{1}{2}$ with $h(n)$ as defined in (7.12). Note that because of the conditions (i) and (ii) the set $\sigma(L) \setminus \sigma_1(L)$ contains only a finite number of eigenvalues. An important remark is that for each $\mu_n \in \sigma_1(L)$ there is exactly one eigenvalue $\lambda_n \in \sigma(L+\Pi)$ for which $|\lambda_n - \mu_n| < \frac{1}{2} i(n)$ and the algebraic multiplicity of this eigenvalue λ_n is equal to 1.

We observe namely that as a consequence of Theorem 5.2 for each $\delta \in [0,1]$ the curve $\Gamma_n = \{\lambda \,|\, |\lambda - \mu_n| = \frac{1}{2} i(n)\}$ separates the spectrum $\sigma(L+\delta\Pi)$ into two parts. From the theory given in Kato (1966, Chap. IV, §3.5, pp. 213–214) it follows that the part of $\sigma(L+\delta\Pi)$ inside Γ_n consists for all $\delta \in [0,1]$ of exactly one point $\lambda_n(\delta)$ which has multiplicity 1 and varies continuously with δ. This implies the remark given above.

Let us now for $\mu_n \in \sigma_1(L)$ consider the difference between the one-dimensional projections $P(\lambda_n)$ and $E(\mu_n)$. We have:

(7.28)

$$
\begin{aligned}
P(\lambda_n)-E(\mu_n) &= -\frac{1}{2\pi i}\oint_{\Gamma_n}\left[(L+\Pi-\zeta)^{-1}-(L-\zeta)^{-1}\right]d\zeta \\
&= \frac{1}{2\pi i}\oint_{\Gamma_n}(L+\Pi-\zeta)^{-1}\Pi(L-\zeta)^{-1}d\zeta \\
&= \frac{1}{2\pi i}\oint_{\Gamma_n}(L-\zeta)^{-1}\Pi(L-\zeta)^{-1}d\zeta \\
&\quad -\frac{1}{2\pi i}\oint_{\Gamma_n}\left\langle (L-\zeta)^{-1}c, \Omega(\zeta)^{-1}P(L-\zeta)^{-1}\Pi(L-\zeta)\cdot\right\rangle d\zeta \\
&= -E(\mu_N)\Pi(L-\mu_n)_*^{-1}-(L-\mu_n)_*^{-1}\Pi E(\mu_n) \\
&\quad -\frac{1}{2\pi i}\oint_{\Gamma_n}\left\langle (L-\zeta)^{-1}c, \Omega(\zeta)^{-1}P(L-\zeta)^{-1}\Pi(L-\zeta)^{-1}\cdot\right\rangle d\zeta.
\end{aligned}
$$

The first and second equalities in (7.28) are obvious. For the third equality in (7.28) we used (3.6). For the final equality in (7.28) we used (7.8), case b and Cauchy's residue theorem. Using analogous tricks as in (7.21)–(7.23) we derive the following norm estimate for $P(\lambda_n)-E(\mu_n)$ from (7.28):

(7.29)
$$
\begin{aligned}
&|||P(\lambda_n)-E(\mu_n)|||_{HD(\alpha)} \\
&\leqq 2\hat{c}(n)\left(\frac{Q^{(n)}}{i(n)}+\frac{Q_2}{d_0(n)}\right)+2\hat{P}^{(n)}\left(\frac{C^{(n)}}{i(n)}+\frac{C_2}{d_0(n)}\right) \\
&\quad +4w_0\|P\|_{\{HD(\alpha)'\}^p}i(n)\left(\frac{C^{(n)}}{i(n)}+\frac{C_2}{d_0(n)}\right)^2\left(\frac{Q^{(n)}}{i(n)}+\frac{Q_2}{d_0(n)}\right)
\end{aligned}
$$

with

$$
\begin{aligned}
\hat{c}^{(n)} &= \sqrt{p}\max_{i=1,\cdots,p}|\langle c_i, \hat{\phi}_n\rangle_{HD(\alpha)}|, \\
\hat{P}^{(n)} &= \sqrt{p}\max_{i=1,\cdots,p}|P_i\hat{\phi}_n|, \\
d_0(n) &= \begin{cases} \mu_1-\mu_{n-1}+\frac{1}{2}i(n) & \text{for } n\geqq 2, \\ \infty & \text{for } n=1. \end{cases}
\end{aligned}
$$

Note that because of (2.1.5) and condition (iii) we have $\Sigma_{n=1}^{\infty} d_0(n)^{-2}<\infty$. If we call the right-hand side in the estimate (7.29) rhs(n) then rhs(n) makes sense for all $n\in\mathbb{N}$ and it is not difficult to verify that $\Sigma_{n=1}^{\infty}\text{rhs}(n)<\infty$. Now define $\sigma_1(L+\Pi)=\sigma(L+\Pi)\cap\{\cup_{\mu_n\in\sigma_1(L)}\{\lambda\,|\,|\lambda-\mu_n|<\frac{1}{2}i(n)\}\}$. It is clear that $\sigma(L+\Pi)\backslash\sigma_1(L+\Pi)$ is finite. For a

subset $\Lambda \subset \sigma(L+\Pi)$ we now decompose:

$$(7.30)\qquad \sum_{\lambda\in\Lambda} P(\lambda) = \sum_{\lambda\in\sigma_1(L+\Pi)\cap\Lambda} P(\lambda) + \sum_{\substack{n\\ \lambda_n\in\sigma_1(L+\Pi)\cap\Lambda}} \big(P(\lambda_n)-E(\mu_n)\big) + \sum_{\substack{n\\ \lambda_n\in\sigma_1(L+\Pi)\cap\Lambda}} E(\mu_n).$$

From (7.30) we deduce

$$(7.31)\qquad \Big|\!\Big|\!\Big| \sum_{\lambda\in\Lambda} P(\lambda) \Big|\!\Big|\!\Big| \leqq \sum_{\lambda\in\sigma(L+\Pi)\setminus\sigma_1(L+\Pi)} |||P(\lambda)||| + \sum_{\substack{n\\ \lambda_n\in\sigma_1(L+\Pi)}} |||P(\lambda_n)-E(\mu_n)||| + \Big|\!\Big|\!\Big| \sum_{\substack{n\\ \lambda_n\in\sigma_1(L+\Pi)\cap\Lambda}} E(\mu_n) \Big|\!\Big|\!\Big|$$

$$\leqq \sum_{\lambda\in\sigma(L+\Pi)\setminus\sigma_1(L+\Pi)} |||P(\lambda)||| + \sum_{n=1}^{\infty} \mathrm{rhs}(n) + 1.$$

Here $|||\ |||$ is a shorthand notation for $|||\ |||_{HD(\alpha)}$.

In the second inequality of (7.31) we used the orthogonality of the projections $E(\mu_n)$ in the sense of $\langle\ ,\ \rangle_{HD(\alpha)}$.

Herewith (7.27) has been verified and the proof is complete. □

Note that in the one-dimensional case with $L = d^2/dx^2 - q$ the condition (7.25) is always satisfied.

This follows from the estimate $|\mu_n + n^2\pi^2/1^2| \leqq \max_{x\in D}|q(x)|$, $1 =$ length of the interval D, which holds for n sufficiently large.

Theorem 7.1 can be seen as a generalization of a theorem given in Schwartz (1954) or Kramer (1957) for the special case that their perturbing operator (i.e. Π here) has a finite-dimensional range. The advantage of the theorem given here is that it is also useful in certain nontrivial cases with $d>1$ whereas the conditions for the theorems given in the above mentioned references are such that these cases are automatically excluded.

For example, consider the case of the Laplace operator on a 2-dimensional rectangular domain $\{(x,y)\,|\,0\leqq x\leqq L, 0\leqq y\leqq \sqrt{p}\cdot L\}$ with $p\in\mathbb{R}\setminus\mathbb{Q}$. For each $a\in\mathbb{R}$ there is at most one pair n,m such that $n^2+pm^2=a$. The conclusion is, that all eigenvalues $\pi^2L^{-2}\{n^2+pm^2\}$ are simple; thus condition (i) in Theorem 7.2 is satisfied. In order to satisfy (7.25) in this example it is sufficient that the coefficients $P_i\hat{\phi}_k$ and $\langle c_j, \hat{\phi}_k\rangle_{HD(\alpha)}$ decrease sufficiently fast for $k\uparrow\infty$ compared with $i(k)$.

8. Some remarkable "resolution of the identity" formulae. Using the results on completeness of eigenprojections as given in §7 it is possible to deduce some remarkable formulae of the "resolution of the identity" type. We shall formulate these formulae without any reference to the control context where they come from. This is done in order to emphasize the general character of these formulae, which makes them interesting also outside the control context.

THEOREM 8.1. *Let L be a linear, uniformly elliptic, formally selfadjoint, 2nd order partial differential operator with coefficients $\in C^\infty(\bar{D})$, $D\subset\mathbb{R}^d$ a bounded domain with a smooth boundary. Let $\sigma(L)$ be the spectrum associated to the Dirichlet problem for L. Suppose that each eigenvalue $\in\sigma(L)$ is simple and number these eigenvalues as $\{\mu_n;\mu\in\mathbb{N}\}$ in such a way that $m>n\Rightarrow\mu_m<\mu_n$.*

Let ϕ and Ψ be two functions $\in HD(\gamma)$ with $\gamma \geqq \max(0, \frac{1}{4}(d-2))$. Suppose that ϕ and Ψ are such that the meromorphic function $\omega(\lambda)=1+\langle \Psi,(L-\lambda)^{-1}\phi\rangle_{L_2(D)}$ has (i) *poles of order* 1 *in all points $\mu_n \in \sigma(L)$* (ii) *only simple zeros.*

Then it is possible to number the zeros of $\omega(\lambda)$ as $\{\mu_n; n\in\mathbb{N}\}$ in such a way that for $n\uparrow\infty$: $\mathrm{Im}(\lambda_n)\downarrow -\infty$, $\mathrm{Re}(\lambda_n)\to 0$. Further we have the following resolutions of the identity in $L_2(D)$:

$$I=\sum_{n\in\mathbb{N}}{}' \left[\frac{d\omega}{d\lambda}(\lambda_n)\right]^{-1}\cdot(L-\lambda_n)^{-1}\phi\cdot\left\langle\left(L-\bar{\lambda}_n\right)^{-1}\Psi,\cdot\right\rangle_{L_2(D)}, \tag{8.1}$$

$$I=\sum_{n\in\mathbb{N}}{}' \left[\frac{dw}{d\lambda}(\lambda_n)\right]^{-1}\cdot(L-\lambda_n)^{-1}\Psi\cdot\left\langle\left(L-\bar{\lambda}_n\right)^{-1}\phi,\cdot\right\rangle_{L_2(D)}. \tag{8.2}$$

Here $\sum'_{n\in\mathbb{N}}$ means that the summation has to be done in a special order: $\sum'_{n\in\mathbb{N}}=\sum_{m=1}^{\infty}(\sum_{k(m)}^{k(m+1)-1}\cdot)$ with $k(1)=1$, $m_1>m_2\Rightarrow k(m_1)>k(m_2)$. The convergence in (8.1)–(8.2) *is meant in the following sense:*

$$\sum_{m=1}^{N}\left(\sum_{kk(m)}^{k(m+1)-1}\cdot\right)\to I \text{ for } N\to\infty \text{ strongly in } L_2(D).$$

In the special case $d=1$ and $L=(d^2/dx^2)-q$ one can replace in (8.1) *and* (8.2) *$\sum'_{n\in\mathbb{N}}$ by the normal summation $\sum_{n\in\mathbb{N}}$.*

The sequences of functions $\{(L-\lambda_n)^{-1}\phi; n\in\mathbb{N}\}$ and $\{(L-\bar{\lambda}_n)^{-1}\Psi; n\in\mathbb{N}\}$ which appear in (8.1) *and* (8.2) *satisfy a biorthogonality condition in $L_2(D)$:*

$$\left\langle\left(L-\bar{\lambda}_k\right)^{-1}\Psi,(L-\lambda_m)^{-1}\phi\right\rangle_{L_2(D)}=0 \quad \text{if } k\neq m. \tag{8.3}$$

Proof of Theorem 8.1. Consider the operator $L+\Pi_1$ with

$$\Pi_1=\phi\langle\Psi,\cdot\rangle_{L_2(D)}. \tag{8.4}$$

Now the theory developed in the preceding sections is applicable to the Dirichlet problem for $L+\Pi$ with $X=Y=L_2(D)$. Note that $\omega(\lambda)$ is the Weinstein–Aronszajn determinant corresponding to $L+\Pi_1$, see §3.

It is now clear that $\sigma(L+\Pi_1)$ coincides with the set of zero's of $\omega(\lambda)$ and that each eigenvalue of $L+\Pi_1$ with Dirichlet b.c. has an algebraic multiplicity $=1$.

As a consequence of Theorem 5.2 they can be numbered in the way of Theorem 8.1.

Using Lemma 7.1a we find that the projection on the one-dimensional eigenspace $N(\lambda_n)$, $\lambda_n\in\sigma(L+\Pi_1)$ is given by:

$$\begin{aligned}P(\lambda_n)&=\left[\frac{d\omega}{d\lambda}(\lambda_n)\right]^{-1}\cdot(L-\lambda_n)^{-1}\phi\cdot\left\langle\Psi,(L-\lambda_n)^{-1}\cdot\right\rangle_{L_2(D)}\\&=\left[\frac{d\omega}{d\lambda}(\lambda_n)\right]^{-1}\cdot(L-\lambda_n)^{-1}\phi\left\langle\left(L-\bar{\lambda}_n\right)^{-1}\Psi,\cdot\right\rangle_{L_2(D)}.\end{aligned} \tag{8.5}$$

The resolution of the identity formula given in (8.1) is now found by applying Theorem 7.2. Note that the condition (7.13) is satisfied since ϕ, $\Psi\in HD(\gamma)$ with $\gamma\geqq\max(0,\frac{1}{4}(d-2))$.

By repeating the arguments above for $L+\Pi_2$ with

$$\Pi_2=\Psi\langle\phi,\cdot\rangle_{L_2(D)} \tag{8.6}$$

the contents of (8.2) are found.

The remark in Theorem 8.1 that in (8.1) $\Sigma'_{n\in\mathbb{N}}$ can be replaced by $\Sigma_{n\in\mathbb{N}}$ in the special case $d=1$, $L=d^2/dx^2q$ follows from Theorem 8.3, especially the remark just below the proof of that theorem.

The biorthogonality relation given in (8.3) follows directly from the following property of the projection $P(\lambda_n)$: $P(\lambda_n)N(\lambda_m)=\{0\}$ for $n\neq m$. Consequently $P(\lambda_n)(L-\lambda_m)^{-1}\phi=0$ for $n\neq m$ and since $[d\omega/d\lambda(\lambda_n)]^{-1}(L-\lambda_n)^{-1}\phi\neq 0$ we have to have (8.3). □

Note that one way to look at Theorem 8.1 is that the theorem, given the operator L, constructs for almost arbitrary ϕ and Ψ biorthogonal sequences of the structure $\{(L-\lambda_n)^{-1}\phi, (L-\bar{\lambda}_n)^{-1}\Psi; n\in\mathbb{N}\}$ which have a completeness property. The possibility of such a construction for almost arbitrary ϕ and Ψ is not at all trivial!

Further we note that the idea of the proof of Theorem 8.1 can be used in many other situations to find interesting resolutions of the identity.

We shall give two other examples.

THEOREM 8.2. *Let L and $\sigma(L)$ satisfy the same conditions as in Theorem* 8.1. *Now suppose that two functions ϕ and $\Psi\in HD(\gamma)$ with $\gamma\geqq\max(0,\frac{1}{4}(d-2))$ are given with the property that $\forall n\in\mathbb{N}$:*

$$\hat{\phi}_n,\hat{\Psi}_n=0 \tag{8.7}$$

with $\hat{\phi}_n=\langle\phi,\phi_n\rangle_{L_2(D)}$, $\hat{\Psi}_n=\langle\Psi,\phi_n\rangle_{L_2(D)}$, ϕ_n as in (2.1), (2.2).

Then the following resolution of the identity holds true in $L_2(D)$:

$$I=\sum_{n\in\mathbb{N}}{}'\big(\phi_n-\alpha_n\hat{\Psi}_n(L-\mu_n)_*^{-1}\phi\big)\big\langle\phi_n-\beta_n\hat{\phi}_n(L-\mu_n)_*^{-1}\Psi,\cdot\big\rangle_{L_2(D)} \tag{8.8}$$

with

$$\alpha_n=\begin{cases}0 & \text{if } \hat{\phi}_n\neq 0,\\ 1 & \text{if } \hat{\phi}_n=0,\end{cases}\qquad \beta_n=\begin{cases}0 & \text{if } \hat{\Psi}_n\neq 0,\\ 1 & \text{if } \hat{\Psi}_n=0.\end{cases}$$

The sequence $\{\phi_n-\alpha_n\hat{\Psi}_n(L-\mu_n)_*^{-1}\phi; n\in\mathbb{N}\}$ are biorthonormal in $L_2(D)$. In the special case $d=1$, $L=d^2/dx^2-q$ we can replace $\Sigma'_{n\in\mathbb{N}}$ by the normal summation $\Sigma_{n\in\mathbb{N}}$.

Proof of Theorem 8.2. In this case we have $\omega(\lambda)=1+\langle\Psi,(L-\lambda)^{-1}\phi\rangle\equiv 1$ on $\mathbb{C}$. For the operator $L+\Pi$ with $\Pi=\phi\langle\Psi,\ \rangle_{L_2(D)}$ we find from the Weinstein–Aronzajn theory that $\sigma(L+\Pi)=\sigma(L)$ and that $\mu_n\in\sigma(L+\Pi)$ has an algebraic multiplicity $=1$. For the eigenprojection on the 1-dimensional eigenspace corresponding to $\mu_n\in\sigma(L+\Pi)$ we find using Lemma 8.1b:

$$\begin{aligned}P(\mu_n)&=E(\mu_n)-(L-\mu_n)_*^{-1}\big\langle E(\mu_n)\Psi,\cdot\big\rangle_{L_2(D)}\\&\quad-E(\mu_n)\phi\big\langle(L-\mu_n)_*^{-1}\Psi,\cdot\big\rangle_{L_2(D)}\\&=E(\mu_n)-(L-\mu_n)_*^{-1}\big\langle E(\mu_n)\Psi,\cdot\big\rangle_{L_2(D)}\quad\text{if }\hat{\phi}_n=0,\\&\quad E(\mu_n)-E(\mu_n)\phi\big\langle(L-\mu_n)_*^{-1}\psi,\cdot\big\rangle_{L_2(D)}\quad\text{if }\hat{\Psi}_n=0\\&=\big(\phi_n-\hat{\Psi}_n(L-\mu_n)_*^{-1}\phi\big)\langle\phi_n,\cdot\rangle_{L_2(D)}\quad\text{if }\hat{\phi}_n=0\\&\quad\phi_n\big\langle\phi_n-\hat{\phi}_n(L-\mu_n)_*^{-1}\Psi,\cdot\big\rangle_{L_2(D)}\quad\text{if }\hat{\Psi}_n=0\\&=\big(\phi_n-\alpha_n\hat{\Psi}_n(L-\mu_n)_*^{-1}\phi\big)\big\langle\phi_n-\beta_n\hat{\Phi}_n(L-\mu_n)^{-1}\Psi,\cdot\big\rangle_{L_2(D)}.\end{aligned} \tag{8.9}$$

Using Theorem 7.2, we then find (8.8). The biorthogonality of the sequences of functions associated to (8.8) follows from the projection properties of the $P(\mu_n)$'s, namely:

$$P(\mu_n)\big(\phi_m - \alpha_m \hat{\Psi}_m (L-\mu_m)_*^{-1}\phi\big) = \begin{cases} 0 & \text{if } m \neq n \\ \phi_n - \alpha_n \Psi_n (L-\mu_n)_*^{-1} & \text{if } m = n. \end{cases}$$

The last statement of Theorem 8.2 is derived analogous to the analogous statement in Theorem 8.1. □

THEOREM 8.3. *Let L and $\sigma(L)$ be as in Theorem* 8.1. *Here we shall suppose that the dimension of D satisfies $d \leqq 3$. Now let ϕ be a function $\in HD(\alpha+2\gamma)$ with $\alpha > \frac{d}{4}$, $\gamma \geqq \min(0, \frac{1}{4}(d-2))$ and let y be a point $\in D$. Suppose that ϕ and y are such that the meromorphic function $\omega(\lambda) = 1 + [(L-\lambda)^{-1}\phi](y)$ has* (i) *poles of order* 1 *in all points of $\sigma(L)$ and* (ii) *only simple zero's. Then it is possible to number the zeros of $\omega(\lambda)$ as $\{\lambda_n, n \in \mathbb{N}\}$ in such a way that for $n \uparrow \infty \Rightarrow \operatorname{Im}\lambda_n \downarrow 0$, $\operatorname{Re}\lambda_n \to 0$.*

The following resolution of the identity formula is valid for $HD(\alpha+\gamma)$:

$$I = \sum_{n\in\mathbb{N}}{}' \left[\frac{d\omega}{d\lambda}(\lambda_n)\right]^{-1} (L-\lambda_n)^{-1}\phi \left\langle G(y,\cdot,\bar{\lambda}_n), \cdot \right\rangle_{L_2(D)} \tag{8.10}$$

with $G(x,\xi;\lambda)$ the Green's kernel of the Dirichlet problem for $L-\lambda$. The sequences of functions $\{(L-\lambda_n)^{-1}\phi; n \in \mathbb{N}\}$ and $\{G(y,\cdot;\bar{\lambda}_n); n \in \mathbb{N}\}$ are biorthogonal in $L_2(D)$.

In the special case $d=1$, $L=(d^2/dx^2)-q$ we can replace $\Sigma'_{n\in\mathbb{N}}$ in (8.10) *by the normal summation $\Sigma_{n\in\mathbb{N}}$.*

Proof of Theorem 8.3. Because of Sobolev's theorem (see Adams (1975)) and the imbeddings given in (2.2.7) we see that for $\alpha > d/4$ the functional δ_y which maps u to $u(y)$ is a well-defined element of $HD(\alpha)'$. $\omega(\lambda)$ is now the Weinstein–Aronszajn determinant corresponding to $L+\Pi$ with $\Pi = \phi\delta_y$. Here we consider Π as an element of $\mathscr{L}(X \to Y)$ with $X = Y = HD(\alpha+\gamma)$.

From here on the proof continues completely analogous to the proof of Theorem 8.1. We now find

$$P(\lambda_n) = \left[\frac{d\omega}{d\lambda}(\lambda_n)\right]^{-1} \cdot (L-\lambda_n)^{-1}\phi \cdot \delta_y (L-\lambda_n)^{-1}. \tag{8.11}$$

Since we required that $d \leqq 3$ the singularity of the Green's kernel $G(\chi,\zeta;\lambda_n)$ is quadratically integrable and we can rewrite

$$\delta_y (L-\lambda_n)^{-1} = \left\langle G(y,\cdot;\bar{\lambda}_n), \cdot \right\rangle_{L_2(D)}. \tag{8.12}$$

Next we apply Theorem 7.1. The condition (7.12)–(7.13) is satisfied since the $C^{(n)}$ corresponding to ϕ and the $Q^{(n)}$ corresponding to δ_y are both $O((\mu_1-\mu_n)^{-\gamma})$.

Thus we are led to (8.10). Further details of the proof are left to the reader. □

9. On the possibility of generalizations. Many generalizations of the results developed in the preceding sections can probably be given. Let us discuss a few of them. It is rather obvious that most of the results can be generalized to feed-back control problems for 2nd order elliptic operators L as in 1.4, but with b.c. of Neumann-type, such that the operator becomes formally selfadjoint. Generalizations to feed-back control of higher order elliptic formally selfadjoint operators subject to Dirichlet or other b.c. that make the operator selfadjoint are certainly also possible. Generalizations of our results to some cases with L as in (1.4) and Dirichlet or other formally selfadjoint b.c. but with a domain D with corners can also be given. Here one should be reminded that

$(L-\lambda)^{-1}$ with $\lambda \notin \sigma(L)$ operates from $L_2(D)$ to $H^2(D)$, but in general even for $f \notin C^\infty(\overline{D})$ the regularity at the corners in the boundary of $(L-\lambda)^{-1}f$ is not better than $H^2(D)$.

Appendix. In this appendix $\subset$ will always be meant in the following sense: "can be considered as a subset, where the canonical injection is bounded". Let us start with some simple observations: $n \in \mathbb{N} \Rightarrow$

$$\text{(A.1)} \qquad CD(n) \supset \left\{u \in C^{2n}(\overline{D}) \,\middle|\, A^k u = 0 \text{ on } \partial D, 0 \leqq k < n\right\},$$
$$C_0 D(n) \supset \left\{u \in C^{2n}(\overline{D}) \,\middle|\, A^k u = 0 \text{ on } \partial D, 0 \leqq k \leqq n\right\}.$$

However the following proposition is less trivial.

PROPOSITION A.1. *For* $\alpha > 0$ *and* $0 \leqq \beta < \alpha$

$$\text{(A.2)} \qquad C_0 D(\alpha) \subset \left\{u \in C^{2\beta}(\overline{D}) \,\middle|\, A^k u = 0 \text{ on } \partial D, 0 \leqq k \leqq \beta\right\}.$$

Proof of Proposition A.1. Using (repeatedly) the relation $C_0 D(\alpha+1) = A^{-1} C_0 D(\alpha)$ and the properties of A^{-1} (see (2.1.7) and (2.2.10)) we see that it is sufficient to show (A.2) for $0 < \alpha < 1$.

For $\alpha = 1$ it is possible to prove (A.2) from the properties of the Green's kernel of A^{-1} (see Ladyzhenskaya and Ural'tseva (1968)) analogous to their calculations on the pp. 110–120.

Further details are left to the reader.

In the case $0 < \alpha < 1$ the following lemma will be useful.

LEMMA A.1. *Let* $\beta \in (0,1)$ *be given. Then*:

a. $\forall \varepsilon \in (0, 1-\beta)\ \exists K(\varepsilon) > 0\ \forall u \in C_0 D(1)$:

$$\text{(A.3)} \qquad \|u\|_{C^{2\beta}(\overline{D})} \leqq K(\varepsilon) \|Au\|^{\beta+\varepsilon}_{C(\overline{D})} \|u\|^{1-\beta-\varepsilon}_{C(\overline{D})}.$$

b. $\forall \varepsilon \in (0, 1-\beta)\ \exists \hat{K}(\varepsilon) > 0\ \forall t \geqq 0\ \forall f \in C_0(\overline{D})$:

$$\text{(A.4)} \qquad \left\|(A+t)^{-2} f\right\|_{C^{2\beta}(\overline{D})} \leqq \hat{K}(\varepsilon)(1+t)^{-2+\beta+\varepsilon} \|f\|_{C(\overline{D})}.$$

Proof of Lemma A.1.

a. We have already shown that $C_0 D(1) \subset C^{2\beta}(\overline{D})$. Because of an "interpolation" argument for $C^{2\beta}(\overline{D})$ in between $C^{2\delta}(\overline{D})$, $\beta < \delta < 1$, and $C(\overline{D})$, (see Miranda (1970)) the following estimate is valid:

$$\|u\|_{C^{2\beta}(\overline{D})} \leqq K_1 \|u\|^{\beta/\delta}_{C^{2\delta}(\overline{D})} \|u\|^{1-\beta/\delta}_{C(\overline{D})}.$$

Using the continuity of the injection associated to $C_0 D(1) \subset C^{2\beta}(\overline{D})$ we find

$$\|u\|_{C^{2\beta}(\overline{D})} \leqq K_2 \|Au\|^{\beta/\delta}_{C(\overline{D})} \|u\|^{1-\beta/\delta}_{C(\overline{D})}.$$

By putting $\beta/\delta = \beta + \varepsilon$ (A.3) is obtained.

b. Since $(A+t)^{-2} f \in C_0 D(1)$ we can use (A.3). This yields

$$\left\|(A+t)^{-2} f\right\|_{C^{2\beta}(\overline{D})} \leqq K \left\|A(A+t)^{-2} f\right\|^{\beta+\varepsilon}_{C(\overline{D})} \left\|(A+t)^{-2} f\right\|^{1-\beta-\varepsilon}_{C(\overline{D})}$$
$$\leqq \hat{K}(1+t)^{-(\beta+\varepsilon)} (1+t)^{-2(1-\beta-\varepsilon)} \|f\|_{C(\overline{D})}.$$

For the latter inequality we have used the formula $A(A+t)^{-2} = (A+t)^{-1} - t(A+t)^{-2}$ and (2.2.3). □

Now we shall show that for $0<\alpha<1$ and $0\leqq\beta<\alpha$ there is a constant $M>0$ such that $\forall f\in C_0(\bar{D})$

$$\|A^{-\alpha}f\|_{C^{2\beta}(\bar{D})}\leqq M\|f\|_{C(\bar{D})}. \tag{A.5}$$

By definition (see Krasnoselskii (1976, p. 281)) we have

$$A^{-\alpha}f=C(\alpha)\int_0^\infty t^{1-\alpha}(A+t)^{-2}f\,dt.$$

Because of (A.4) the integrand satisfies

$$\left\|t^{1-\alpha}(A+t)^{-2}f\right\|_{C^{2\beta}(\bar{D})}\leqq\hat{K}(\varepsilon)t^{-1-(\alpha-\beta-\varepsilon)}$$

where $\varepsilon>0$ is still a free parameter. By choosing $\varepsilon=\frac{1}{2}(\alpha-\beta)$ we see that the integral makes sense in $C^{2\beta}(\bar{D})$ and that (A.5) holds true.

From (A.5) we can conclude $C_0D(\alpha)\subset C^{2\beta}(\bar{D})\cap C_0(\bar{D})$ and this completes the proof of (A.2). □

Our following step will be to prove the next proposition.

PROPOSITION A.2. *For $\alpha>0$ and $\beta>\alpha$:*

$$C_0D(\alpha)\supset\left\{u\in C^{2\beta}(\bar{D})\,\middle|\,A^k u=0 \text{ on } \partial D, 0\leqq k\leqq\alpha\right\}. \tag{A.6}$$

Proof. It is again sufficient to show (A.6) for $0<\alpha\leqq1$. For $\alpha=1$ it is clear that (A.6) is contained in (A.1).

For $0<\alpha<1$ the following lemma will be useful.

LEMMA A.2. *Let $\gamma\in(0,1)$ be given. There exists a constant $K(\gamma)>0$ such that $\forall t\geqq0$ $\forall f\in C^{2\gamma}(\bar{D})$:*

$$\left\|A(A+t)^{-1}f\right\|_{C(\bar{D})}\leqq K(\gamma)(1+t)^{-\gamma}\|f\|_{C^{2\gamma}(\bar{D})}. \tag{A.7}$$

Proof of Lemma A.2. Let us introduce the notation $u=A(A+t)^{-1}f$. We shall decompose u in the following way:

$$\begin{aligned} &u=(f-\tilde{f})+(\tilde{f}-\tilde{w})+(\tilde{w}-w),\\ &w=t(A+t)^{-1}f,\qquad \tilde{w}=t(A+t)^{-1}\tilde{f}. \end{aligned} \tag{A.8}$$

The function $\tilde{f}$ will be an approximation of f which we construct from f by a "smoothing" process:

$$\tilde{f}(x)=\int_{\mathbb{R}^n}f^{\text{ext}}(y)\phi^\varepsilon(x-y)\,dy.$$

Here f^{ext} is an extension of f from $\bar{D}$ to the whole of $\mathbb{R}^n$ such that

$$\|f^{\text{ext}}\|_{C^{2\gamma}(\mathbb{R}^n)}\leqq2\|f\|_{C^{2\gamma}(\bar{D})}.$$

Further: $\phi^\varepsilon(x)=\phi(\varepsilon x)/\varepsilon^n$ with ϕ a spherical symmetric, positive, ∞-differentiable function which satisfies:

$$\begin{aligned} &\phi(x)=0\quad\text{for } \|x\|\geqq1,\\ &\int_{\mathbb{R}^n}\phi(x)\,dx=1. \end{aligned}$$

The parameter ε is taken equal to $(1+t)^{-1/2}$.

We shall now estimate the norms of the three terms in the decomposition (A.6):

(i) It is not difficult to verify that

$$\|f-\tilde{f}\|_{C(\bar{D})} \leqq K_1 \varepsilon^{2\gamma} \|f\|_{C^{2\gamma}(\bar{D})}.$$

(ii) Using (2.3.3) we find that

$$\|w-\tilde{w}\|_{C(\bar{D})} \leqq t(1+t)^{-1} \|f-\tilde{f}\|_{C(\bar{D})} \leqq K_1 \varepsilon^{2\gamma} \|f\|_{C^{2\gamma}(\bar{D})}.$$

(iii) Since $\tilde{f}-\tilde{w} = A(A+t)^{-1}\tilde{f}$ we obtain by applying (2.3.3)

$$\|\tilde{f}-\tilde{w}\|_{C(\bar{D})} \leqq (1+t)^{-1} \|A\tilde{f}\|_{C(\bar{D})}.$$

It is easy to see that

$$\|A\tilde{f}\|_{C(\bar{D})} \leqq K_2 \varepsilon^{2\gamma-2} \|f\|_{C^{2\gamma}(\bar{D})}.$$

By applying the triangle inequality to (A.8) and next substituting the estimates given in (i), (ii), (iii) the desired result (A.7) is found. □

Let us continue the proof of (A.6). Of course we can now continue ourselves to the situation $0 < \alpha < \beta < 1$. For $u \in C_0 D(1)$ we have

$$A^{\alpha} u = C(\alpha) \int_0^{\infty} t^{\alpha} (A+t)^{-1} \cdot A(A+t)^{-1} u \, dt$$

and

$$\|A^{\alpha} u\|_{C(\bar{D})} \leqq C(\alpha) \int_0^{\infty} t^{\alpha} (1+t)^{-1} \|A(A+t)^{-1} u\|_{C(\bar{D})} dt.$$

Because of (A.2) the function $u \in C_0 D(1)$ is certainly an element of $C^{2\beta}(\bar{D})$ and Lemma A.2 can be used to estimate $\|A(A+t)^{-1}u\|_{C(\bar{D})}$. In this way we obtain

$$\|A^{\alpha} u\|_{C(\bar{D})} \leqq K \int_0^{\infty} t^{\alpha} (1+t)^{-(\beta+1)} dt \cdot \|u\|_{C^{2\beta}(\bar{D})}. \tag{A.9}$$

The estimate in (A.9) is not only valid for $u \in C_0 D(1)$, but for all $u \in C^{2\beta}(\bar{D}) \cap C_0(\bar{D})$. This follows from the fact that $C_0 D(1)$ is dense in $C^{2\beta}(\bar{D}) \cap C_0(\bar{D})$ (see A.1!). Herewith the proof is complete. □

It is obvious that (2.2.8) is implied by Propositions A.1 and A.2. We shall continue this appendix by proving (2.2.9).

PROPOSITION A.3. *For* $\alpha > 0$ *and* $0 \leqq \beta < \alpha$:

$$CD(\alpha) \subset C_0 D(\beta). \tag{A.10}$$

Proof of Proposition A.3. It is again sufficient to show that (A.10) holds true for $0 < \alpha \leqq 1$. An argument analogous to the one given in the proof of Proposition A.1 gives the result

$$CD(1) \subset C^{2\beta}(\bar{D}) \cap C_0(\bar{D}) \quad \text{for } 0 \leqq \beta < 1.$$

Because of Proposition A.2 this implies

$$CD(1) \subset C_0 D(\beta) \quad \text{for } 0 \leqq \beta < 1.$$

Hence if we put $\alpha = 1-\delta$, $\beta = \alpha - \varepsilon$, $0 < \varepsilon \leqq 1-\delta$ then it is also true that

$$CD(\alpha) = A^{\delta} CD(1) \subset A^{\delta} C_0 D(1-\varepsilon) = C_0 D(1-\varepsilon-\delta) = C_0 D(\beta). \qquad \square$$

Let us conclude this appendix with the following result.

PROPOSITION A.4. *For $\alpha > 0$, $\alpha \notin N$ and β such that $\alpha \leqq \beta \leqq [\alpha]+1$ it holds true that*

$$CD(\beta) \text{ is dense in } CD(\alpha). \tag{A.11}$$

Proof. Let u be an element of $CD(\alpha)$. Then there exists a sequence $\{f_n; n \in \mathbb{N}\}$ in $C^{\infty}(\overline{D})$ such that $f_n \to A^{\alpha} u$ for $n \to \infty$ in $C(\overline{D})$. Using regularity theory and (A.2), (A.3), we see that $A^{-\alpha} f_n \in \{w \in C^{\infty}(\overline{D}) \mid A^k w = 0 \text{ on } \partial D, 0 \leqq k \leqq [\alpha]\}$. Consequently $A^{-\alpha} f_n \in CD([\alpha]+1)$ and $A^{-\alpha} f_n \to u$ for $n \to \infty$ in $CD(\alpha)$. This implies (A.11). $\square$

REFERENCES

R. A. ADAMS (1975), *Sobolev Spaces*, Academic Press, London.

S. AGMON, A. DOUGLIS AND N. NIRENBERG (1959), *Estimates near the boundary for solutions of elliptic* PDE *satisfying general* BC, I, Comm. Pure Appl. Math., 12, pp. 623–727.

N. ARONSZAJN AND A. WEINSTEIN (1942), *On a unified theory of eigenvalues of plates and membranes*, Amer. J. Math., 64, pp. 623–645.

R. F. CURTAIN AND A. J. PRITCHARD (1978), *Infinite dimensional linear systems theory*, Lecture Notes in Control and Information Sciences, 8, Springer Berlin.

N. DUNFORD AND J. T. SCHWARTZ (1963), *Linear Operators, Part* II: *Spectral Theory*, Interscience, New York.

S. D. EIDEL'MAN (1969), *Parabolic Systems*, North-Holland, Amsterdam.

A. FRIEDMAN (1969), *Partial Differential Equations*, Holt, Rinehart and Winston, New York.

P. R. GARABEDIAN, (1964), *Partial Differential Equations*, John Wiley, New York.

I. C. GOHBERG AND M. G. KREIN (1969), *Introduction to the theory of linear non-selfadjoint operators*, Transl. AMS, 18.

J. HALE (1971), *Functional Differential Equations*, Springer, Berlin.

A. VAN HARTEN AND J. M. SCHUMACHER (1977a), *Some boundary value problems for a class of 2nd order elliptic partial functional differential equations arising in feed-back control theory*, Report 69, V. U. Amsterdam.

_____ (1977b), *On a class of partial functional differential equations arising in feed-back control theory*, Proc. 3rd Scheveningen Conference on Differential Equations, North-Holland, Math. Studies, Vol. 31, North-Holland, Amsterdam, pp. 161–179.

_____ (1978), *Feed-back control of systems with distributed parameters using a finite number of observators and control inputs*, Séminaires IRIA, 1978, pp. 257–272.

_____ (1980), *Well-posedness of some evolution problems in the theory of automatic feed-back control for systems with distributed parameters*, this Journal, 18, pp. 391–420.

A. VAN HARTEN (1979), *On the spectral properties of a class of elliptic* FDE *arising in feed-back control theory for diffusion processes*, Preprint No. 130, M. I., R. U. Utrecht.

T. KATO (1966), *Perturbation Theory for Linear Operators*, Springer, Berlin.

H. P. KRAMER (1957), *Perturbation of differential operators*, Pacific J. Math., 7, pp. 1405–1435.

M. A. KRASNOSELSKII (1976), *Integral Operators in Spaces of Summable Functions*, Noordhoff, Leiden.

O. A. LADYZHENSKAYA, V. A. SOLONNIKOV AND N. N. URAL'TSEVA (1967), *Linear and quasi-linear equations of parabolic type*, Trans. Math. Monographs American Mathematical Society, Providence, RI.

O. A. LADYZHENSKAYA AND N. N. URAL'TSEVA (1968), *Equations aux derivées partielles de type elliptique*, Dunod, Paris.

J. L. LIONS AND E. MAGENES (1972), *Non-homogeneous Boundary Value Problems and Applications, Vol.* I, Springer, Berlin.

S. MINAKSHISUNDARAM AND A. PLEYEL (1949), *Some properties of the eigenfunctions of the Laplace operator on Riemannian manifolds*, Can. J. Math., 2.

C. MIRANDA (1970), *Partial Differential Equations of Elliptic Type*, Springer-Verlag, Berlin.

M. H. PROTTER AND H. F. WEINBERGER (1967), *Maximum Principles in Differential Equations*, Prentice-Hall, Englewood Cliffs, NJ.

J. SCHWARTZ (1954), *Perturbations of spectral operators and applications*, I: *Bounded perturbations*, Pacific J. Math., 4, pp. 415–458.

H. B. STEWART (1974), *Generation of analytic semigroups by strongly elliptic operators*, Trans. Amer. Math. Soc., 199, pp. 141–162.

_____ (1980), *Generation of analytic semigroups by strongly elliptic operators, under general boundary conditions*, Trans. Amer. Math. Soc., 259, pp. 299–310.

TRÈVES (1975), *Basic Linear Partial Differential Equations*, Academic Press, New York.

R. TRIGGIANI (1980), *Boundary feed-back stabilizability of parabolic equations*, Appl. Math. J. Opt, 6, pp. 201–220.

J. H. WILKINSON (1965), *The Algebraic Eigenvalue Problem*, Oxford Univ. Press, Oxford.

SIAM J. MATH. ANAL.
Vol. 17, No. 2, March 1986

ON THE CONSTRUCTION OF SERIES SOLUTIONS TO THE FIRST BIHARMONIC BOUNDARY VALUE PROBLEM ON A RECTANGLE*

CHARLES V. COFFMAN†

Abstract. This paper is concerned with the problem of finding a biharmonic function u on a rectangle Ω which vanishes on $\partial\Omega$ and has a preassigned normal derivative there. The solution u is developed in a series; however the family of functions that enters into this series development is not orthogonal and thus the determination of the coefficients requires the solution of an infinite system of linear equations. In implementation of this method of course the coefficients are obtained by solving a truncation of the original infinite system. The bulk of the paper is devoted to the estimation of the error that results from this truncation.

AMS(MOS) subject classifications. Primary 35J40, 31A30, 65N99

Key words. clamped rectangular plate, series solution

1. Introduction. The difficulty in solving the first biharmonic boundary value problem on a rectangle Ω may be said to reside in the particular case

$$\Delta^2 u = 0 \quad \text{in } \Omega, \tag{1.1}$$

$$u = 0, \quad \frac{\partial u}{\partial \nu} = g \quad \text{on } \partial\Omega, \tag{1.2}$$

(ν denotes the interior normal). That is, the first boundary value problem with general data can be decomposed into problems solvable by separation of variables and a problem of the form (1.1), (1.2).

In this paper we discuss the representation of solutions to (1.1), (1.2) as infinite series involving trigonometric and hyperbolic functions. To determine the coefficients in this expansion, one must solve an infinite system of linear equations. The main purpose of this paper is to obtain a priori estimates for the error that results from truncation of this infinite system.

The methods employed here have a long history in the literature of both pure and applied mathematics. Two principal ideas are involved: first, the reduction of the general problem to the particular problem (1.1), (1.2); and second, the choice of the appropriate set of functions in which to expand the solution to the latter problem.

Except for a difference in choice of coordinates, the procedure described below, as it is applied to the problem of the clamped plate in §7, is precisely the same as that presented by Timoshenko in [13] (see also [12]). The basic ideas involved trace back to the beginning of the century (for references see [12]). The reduction to (1.1), (1.2) is accomplished by solving the problem of the simply supported plate, or by finding some other solution w to (7.1) that vanishes on $\partial\Omega$ and then solving (1.1), (1.2), with $g = \partial w/\partial \nu$, for the biharmonic function which when added to w gives the solution for the clamped plate. This procedure is easily seen to be equivalent to the method of Zaremba [1], [2], [4], [6], or the closely related method of the "nonharmonic residue" of Rafal'son [11] (see also [10]).

*Received by the editors March 29, 1983, and in final revised form October 19, 1984. This research was supported by the National Science Foundation under grant MCS 80-02851.

† Department of Mathematics, Carnegie-Mellon University, Pittsburgh, Pennsylvania 15213.

The expansion used below (formula (3.10)) can be traced back at least to Lauricella [8] (however the procedure given there for determination of the coefficients is in error). Prior to [2] no theoretical justification had been given for this representation and of course neither had convergence estimates been given. The justification that is given in [2] reduces to the demonstration that the angle θ, with respect to the inner product (3.7), between the spaces generated respectively by the functions (3.1), $j=0,1$, $n=1,2,\cdots$, and (3.2), $j=0,1$, $n=1,2,\cdots$ is positive. The a priori estimates there are in terms of $\cos\theta$; these are based on results of [1] where the procedure in [2] was already outlined. As indicated in [2] these estimates are primarily of theoretical interest and are too pessimistic to be of practical value. Here we obtain a priori estimates on the error involved in the truncation of the infinite system that must be solved and which, while they also may be excessively pessimistic in some cases, are nevertheless of practical value. Specifically, we show that under rather mild assumptions on the smoothness of the solution the error in a given coefficient is $O(N^{-\sigma})$ for any σ with $0<\sigma<3$, where N is the number of equations in the truncated system.

2. Preliminaries. We shall assume that the boundary data g in (1.1), (1.2) is sufficiently smooth that the solution u will have its Laplacian in $L^2(\Omega)$. Accordingly, we define $B_0(\Omega)$ to be the Hilbert space that consists of real-valued functions u that are continuous on $\bar{\Omega}$, biharmonic on Ω, vanish on $\partial\Omega$ and have Laplacian in $L^2(\Omega)$; the inner product on $B_0(G)$ is

$$\langle u,v\rangle=\int\int_{\Omega}\Delta u\Delta v\,dx\,dy. \tag{2.1}$$

Let $G(x,y;\xi,\eta)$ denote the harmonic Green's function for Ω. Then a function u defined on Ω belongs to $B_0(\Omega)$ if and only if

$$u(x,y)=\int\int_{\Omega}G(x,y;\xi,\eta)f(\xi,\eta)\,d\xi\,d\eta, \tag{2.2}$$

where f is harmonic and square-integrable on Ω; in this case of course

$$-\Delta u=f.$$

Since the L^2-harmonic functions form a closed subspace of $L^2(\Omega)$ the completeness of $B_0(\Omega)$ follows from the representation (2.2).

The second boundary condition in (1.2) is assumed to hold in the following weak sense

$$\int\int_{\Omega}\varphi\Delta u\,dx\,dy=-\int_{\partial\Omega}\varphi g\,ds, \tag{2.3}$$

for every φ that is harmonic on some neighborhood of $\bar{\Omega}$.

We specify Ω to be of the form

$$\Omega=\{(x,y):0<x<a,0<y<b\}. \tag{2.4}$$

We then express g in terms of sine series as follows

$$\begin{aligned} g(x,\eta)&=\sum_{n=1}^{\infty}\alpha_n(\eta)\sin\frac{n\pi x}{a}, \qquad \eta=0,b,\\ g(\xi,y)&=\sum_{n=1}^{\infty}\beta_n(\xi)\sin\frac{n\pi y}{b}, \qquad \xi=0,a, \end{aligned} \tag{2.5}$$

where

$$\alpha_n(\eta)=\frac{2}{a}\int_0^a g(x,\eta)\sin\frac{n\pi x}{a}\,dx,$$
$$\beta_n(\xi)=\frac{2}{b}\int_0^b g(\xi,y)\sin\frac{n\pi y}{b}\,dy. \tag{2.6}$$

We shall say that

$$g\in H^k(\partial\Omega)$$

if and only if

$$\sum_{n=1}^{\infty} n^{2k}\left(\alpha_n^2(0)+\alpha_n^2(b)+\beta_n^2(0)+\beta_n^2(a)\right)<\infty. \tag{2.7}$$

By expanding a function $u\in B_0(\Omega)$ in a double sine series on Ω and computing $\partial u/\partial\nu$ therefrom, one easily sees that:

(*) *A necessary condition for* (1.1), (1.2) *to admit a solution* $u\in B_0(\Omega)$ *is that* $g\in H^{1/2}(\partial\Omega)$.

The converse is also true.

PROPOSITION 2.1. *The problem* (1.1), (1.2) *has a solution* $u\in B_0(\Omega)$ *if and only if* $g\in H^{1/2}(\partial\Omega)$.

The proof is deferred to the appendix.

It readily follows from assertion (*) and the weak formulation (2.3) of (1.2) that, for a given g, the problem (1.1), (1.2), has at most one solution in $B_0(\Omega)$.

3. Series representation of the solution. We define the functions $\{\varphi_n^j\}$, $\{\psi_n^j\}$, $j=0,1$ as follows

$$\varphi_n^j(x,y)=\frac{c_n^j}{\sqrt{2n\pi}}\sin\frac{n\pi x}{a}\left[(b-y)\sinh\frac{n\pi y}{a}+(-1)^j y\sinh\frac{n\pi}{a}(b-y)\right], \tag{3.1}$$

$$\psi_n^j(x,y)=\frac{d_n^j}{\sqrt{2n\pi}}\sin\frac{n\pi y}{b}\left[(a-x)\sinh\frac{n\pi x}{b}+(-1)^j x\sinh\frac{n\pi}{b}(a-x)\right], \tag{3.2}$$

where

$$c_n^j=\left(\left(\sinh\frac{n\pi b}{a}+(-1)^j\frac{n\pi ub}{a}\right)\left(\cosh\frac{n\pi b}{a}+(-1)^j\right)\right)^{-1/2}, \tag{3.3}$$

$$d_n^j=\left(\left(\sinh\frac{n\pi a}{b}+(-1)^j\frac{n\pi a}{b}\right)\left(\cosh\frac{n\pi a}{b}+(-1)^j\right)\right)^{-1/2}. \tag{3.4}$$

We then have

$$\Delta\varphi_n^j(x,y)=-\sqrt{2n\pi}\,\frac{c_n^j}{a}\sin\frac{n\pi x}{a}\left[\cosh\frac{n\pi y}{a}+(-1)^j\cosh\frac{n\pi}{a}(b-y)\right], \tag{3.5}$$

$$\Delta\psi_n^j(x,y)=-\sqrt{2n\pi}\,\frac{d_n^j}{b}\sin\frac{n\pi y}{b}\left[\cosh\frac{n\pi x}{b}+(-1)^j\cosh\frac{n\pi}{b}(a-x)\right], \tag{3.6}$$

and

$$\int_0^b \int_0^a \left|\Delta\varphi_n^j(x,y)\right|^2 dx\,dy = \int_0^b \int_0^a \left|\Delta\psi_n^j(x,y)\right|^2 dx\,dy = 1,$$

for $j=0,1, n=1,2,\cdots$.

It is clear that the sets $\{\varphi_n^j: j=0,1; n=1,2,\cdots\}$ and $\{\psi_n^j: j=0,1; n=1,2,\cdots\}$ are orthonormal with respect to the inner product

$$\langle u,v\rangle = \int_0^b \int_0^a \Delta u \Delta v\,dx\,dy. \tag{3.7}$$

Two terms φ_n^j and ψ_m^i will have nonzero inner product only if their parities with respect to the lines $x=a/2$ and $y=b/2$ are the same; this necessary condition for φ_n^j and ψ_m^i to have nonzero inner product can be written

$$(-1)^{n+i} = (-1)^{m+j} = -1. \tag{3.8}$$

(*) *If* (3.8) *holds, then*

(3.9)

$$\langle \varphi_n^j, \psi_m^i\rangle = \frac{8a^2b^2(mn)^{3/2}}{\pi(m^2a^2+n^2b^2)^2} c_n^j d_m^i \left[1+(-1)^{m+1}\cosh\frac{n\pi b}{a}\right]\left[1+(-1)^{n+1}\cosh\frac{m\pi a}{b}\right];$$

otherwise

$$\langle \varphi_n^j, \psi_m^i\rangle = 0.$$

Suppose now that u is the solution to (1.1), (1.2), then by Green's formula, (3.5) and (2.5)

$$\begin{aligned}\langle \varphi_n^j, u\rangle &= -\int_0^a \left(\Delta\varphi_n^j(x,b)g(x,b) + \Delta\varphi_n^j(x,0)g(x,0)\right)dx \\ &= \sqrt{\frac{n\pi}{2}}\, c_n^j\left(\cosh\frac{n\pi b}{a} + (-1)^j\right)\left(\alpha_n(b) + (-1)^j\alpha_n(0)\right),\end{aligned}$$

and similarly

$$\langle \psi_n^j, u\rangle = \sqrt{\frac{n\pi}{2}}\, d_n^j\left(\cosh\frac{n\pi a}{b} + (-1)^j\right)\left(\beta_n(a) + (-1)^j\beta_n(0)\right).$$

We put

$$\begin{aligned}\bar{\alpha}_{n,j} &= \sqrt{\frac{n\pi}{2}}\, c_n^j\left(\cosh\frac{n\pi b}{a} + (-1)^j\right)\left(\alpha_n(b) + (-1)^j\alpha_n(0)\right), \\ \bar{\beta}_{n,j} &= \sqrt{\frac{n\pi}{2}}\, d_n^j\left(\cosh\frac{n\pi a}{b} + (-1)^j\right)\left(\beta_n(a) + (-1)^j\beta_n(0)\right),\end{aligned}$$

and note that, in view of (3.3), (3.4),

$$\begin{aligned}\lim_{n\to\infty} c_n^j \cosh\frac{n\pi b}{a} &= 1, \\ \lim_{n\to\infty} d_n^j \cosh\frac{n\pi a}{b} &= 1,\end{aligned}$$

$j=0,1$, so that $g\in H^k(\partial\Omega)$ if and only if

$$\sum_{n=1}^{\infty} n^{2k-1}\bar{\alpha}_{n,j}^2, \quad \sum_{n=1}^{\infty} n^{2k-1}\bar{\beta}_{n,j}^2<\infty, \qquad j=0,1.$$

We seek to represent the solution u to (1.1), (1.2) in the form

$$u(x,y)=\sum_{j=0}^{1}\sum_{n=1}^{\infty}\left(A_n^j\varphi_n^j(x,y)+B_n^j\psi_n^j(x,y)\right), \tag{3.10}$$

where

$$\sum_{j=0}^{1}\sum_{n=1}^{\infty}\left(\left|A_n^j\right|^2+\left|B_n^j\right|^2\right)<\infty. \tag{3.11}$$

When u is so represented, the coefficients must satisfy the infinite system of linear equations

$$\begin{aligned} &A_{2n-1+i}^j+\sum_{k=1}^{\infty} B_{2k-1+j}^i\left\langle \varphi_{2n-1+i}^j,\psi_{2k-1+j}^i\right\rangle=\bar{\alpha}_{2n-1+i,j},\\ &B_{2n-1+j}^i+\sum_{k=1}^{\infty} A_{2k-1+i}^j\left\langle \psi_{2n-1+j}^i,\varphi_{2k-1+i}^j\right\rangle=\bar{\beta}_{2n-1+j,i},\\ &i,j=0,1, \qquad n=1,2,\cdots. \end{aligned} \tag{3.12}$$

The above computations yield immediately the following.

THEOREM 3.1. *Let u_N denote the projection of u onto*

$$\operatorname{sp}\{\varphi_n^i,\psi_n^i: i=0,1; n=1,2,\cdots,2N\}$$

with respect to the inner product (3.7). *Then*

$$u_N(x,y)=\sum_{j=0}^{1}\sum_{n=1}^{2N}\left(\bar{A}_n^j\varphi_n^j(x,y)+\bar{B}_n^j\psi_n^j(x,y)\right) \tag{3.13}$$

where $\{\bar{A}_n^j,\bar{B}_n^j\}$ satisfy

$$\begin{aligned} &\bar{A}_{2n-1+i}^j+\sum_{k=1}^{N}\bar{B}_{2k-1+j}^i\left\langle \varphi_{2n-1+i}^j,\psi_{2k-1+j}^i\right\rangle=\bar{\alpha}_{2n-1+i,j},\\ &\bar{B}_{2n-1+j}^i+\sum_{k=1}^{N}\bar{A}_{2k-1+i}^j\left\langle \psi_{2n-1+j}^i,\varphi_{2n-1}^j\right\rangle=\bar{\beta}_{2n-1+j,i},\\ &i,j=0,1, \qquad n=1,2,\cdots,N. \end{aligned} \tag{3.14}$$

Proof. If u_N is given by (3.13) and (3.14) holds, then

$$\left\langle u-u_N,\varphi_n^j\right\rangle=\left\langle u-u_N,\psi_n^j\right\rangle=0,$$

for $j=0,1, n=1,2,\cdots,2N$.

4. Row and column estimates. In this section we will derive row and column decay estimates for the inverse of the matrix that appears on the left in (3.12). These will be used in the following sections to estimate the error that results from truncation of

(3.12). The system (3.12) is of the form

$$\begin{aligned} A+\Gamma B&=\alpha,\\ \Gamma^*A+B&=\beta; \end{aligned} \tag{4.1}$$

moreover, it decomposes into four subsystems (corresponding to the four choices of the pair (i,j)) all of the form (4.1). If the matrices $I-\Gamma^*\Gamma$, $I-\Gamma\Gamma^*$ are invertible, then the solution to (4.1) is

$$\begin{aligned} A&=(I-\Gamma\Gamma^*)^{-1}(\alpha-\Gamma\beta),\\ B&=(I-\Gamma^*\Gamma)^{-1}(\beta-\Gamma^*\alpha). \end{aligned}$$

In what follows, we shall assume (4.1) to denote one of the subsystems of (3.12), as indicated above. It then follows from the results of [2], (see Appendix), that

$$\|\Gamma\|<1 \tag{4.2}$$

where, here and throughout, $\|\Gamma\|$ denotes the norm of Γ as an operator on l^2. It follows that the series

$$\Gamma\Gamma^*+(\Gamma\Gamma^*)^2+\cdots+(\Gamma\Gamma^*)^n+\cdots,\quad \Gamma^*\Gamma+(\Gamma^*\Gamma)^2+\cdots+(\Gamma^*\Gamma)^n+\cdots, \tag{4.3}$$

are convergent with respect to the norm $\|\cdots\|$. Here we will be concerned with the convergence of these series with respect to norms of the form

$$\|\Gamma\|_\delta=\sup_{m,n}|\gamma_{mn}|\max\{m^{\delta-1/2}n^{3/2-\delta},m^{3/2-\delta}n^{\delta-1/2}\}, \tag{4.4}$$

where $0<\delta\leqq\frac{1}{2}$.

LEMMA 4.1. *Let* $0\leqq\delta\leqq\frac{1}{2}$, *let* $\Gamma=\{\gamma_{mn}\}$ *satisfy*

$$\overline{\lim_{m\to\infty}}\, m^{(3/2)-\delta}\sum_{n=1}^{\infty}|\gamma_{mn}|n^{\delta-3/2}<\rho \tag{4.5}$$

and let the matrix $\Lambda=\{\lambda_{mn}\}$ *satisfy the column decay inequalities*

$$|\lambda_{mn}|\leqq c_0 m^{\delta-3/2}n^{1/2-\delta},\qquad m,n=1,2,\cdots. \tag{4.6}$$

Then the product $\Lambda'=\Gamma\Lambda$ *satisfies the column decay inequalities*

$$|\lambda'_{mn}|\leqq\left(\max\{C\|\Lambda\|,\rho c_0\}\right)m^{\delta-3/2}n^{1/2-\delta}, \tag{4.7}$$

where C *is a constant that depends only on* Γ.

Proof. Choose N so that

$$m^{3/2-\delta}\sum_{k=1}^{\infty}|\gamma_{mk}|k^{\delta-3/2}<\rho \tag{4.8}$$

for $m\geqq N$. Upon combining (4.6) and (4.8) we then have

$$|\lambda'_{mn}|=\left|\sum_{k=1}^{\infty}\gamma_{mk}\lambda_{kn}\right|\leqq\rho c_0 m^{\delta-3/2}n^{1/2-\delta}, \tag{4.9}$$

for $m \geqq N$. For $m \leqq N$,

$$|\lambda'_{mn}| \leqq \left(\sum_{k=1}^{\infty} |\gamma_{mk}|^2 \right)^{1/2} \left(\sum_{k=1}^{\infty} |\lambda_{kn}|^2 \right)^{1/2} \tag{4.10}$$

$$\leqq (N^{3/2-\delta}\|\Gamma\|)\|\Lambda\| m^{\delta-3/2} n^{1/2-\delta}.$$

The inequality (4.7) thus holds with $C = N^{(3/2)-\delta}\|\Gamma\|$.

LEMMA 4.2. *For $i,j = 0,1$, $0 \leqq \delta \leqq \frac{1}{2}$, we have*

$$\lim_{m\to\infty} \left[m^{3/2-\delta} \sum_{k=1}^{\infty} \left| \left\langle \varphi^j_{2m-1+i}, \psi^i_{2k-1+j} \right\rangle \right| k^{\delta-3/2} \right] = \frac{4}{\pi} \left(\frac{a}{b} \right)^{1-\delta} \int_0^{\infty} \frac{s^{\delta}\, ds}{(1+s^2)^2},$$

$$\lim_{m\to\infty} \left[m^{3/2-\delta} \sum_{k=1}^{\infty} \left| \left\langle \psi^j_{2n-1+i}, \varphi^i_{2k-1+j} \right\rangle \right| k^{\delta-3/2} \right] = \frac{4}{\pi} \left(\frac{b}{a} \right)^{1-\delta} \int_0^{\infty} \frac{s^{\delta}\, ds}{(1+s^2)^2}.$$

Proof. Let i,j be fixed and put

$$s(k,m) = (2k-1+j)b/(2m-1+i)a,$$

and

$$\Delta_k x(k,m) = s(k+1,m) - s(k,m) = 2b/(2m-1+i)a.$$

Then it follows from (3.3), (3.4) and (3.9) that

$$m^{3/2-\delta} \sum_{k=1}^{\infty} \left| \left\langle \varphi^j_{2m-1+i}, \psi^i_{2k-1+j} \right\rangle \right| k^{\delta-3/2}$$

$$= \frac{4}{\pi} \left(\frac{b}{a} \right)^{\delta} (1+f(m)) \sum_{k=1}^{\infty} \frac{[s(k,m)]^{\delta}(1+g(k))\Delta_k s(k,m)}{\left(1+(s(k,m))^2\right)^2},$$

where

$$\lim_{m\to\infty} f(m) = \lim_{k\to\infty} g(k) = 0.$$

The first of the asserted formulas follows readily; the proof of the second is exactly the same.

LEMMA 4.3. *Let (i,j) be fixed $(i,j = 0,1)$ and let $\Gamma = \{\gamma_{mk}\}$ where*

$$\gamma_{mk} = \left\langle \varphi^j_{2m-1+i}, \psi^i_{2k-1+j} \right\rangle. \tag{4.11}$$

Let $0 \leqq \delta \leqq \frac{1}{2}$ and let

$$\|\Gamma\|, \frac{4}{\pi} \int_0^{\infty} \frac{s^{\delta}\, ds}{(1+s^2)^2} < \kappa. \tag{4.12}$$

Then there exists a constant C', depending only on Γ, κ, δ, such that

$$\|(\Gamma\Gamma^*)^n\|_\delta, \|(\Gamma^*\Gamma)^n\|_\delta \leqq C'\kappa^{2n}, \qquad n=1,2,\cdots. \tag{4.13}$$

Proof. Suppose that the inequality (4.13) holds for 1 through n and suppose moreover that

$$C'\left(\frac{a}{b}\right)^{1-\delta}, C'\left(\frac{b}{a}\right)^{1-\delta} \geqq \kappa^{-1}C, \tag{4.14}$$

where C is the constant in Lemma 4.1. It follows then from (4.11), the definition (4.4) and Lemmas 4.1 and 4.2 that $\Gamma^*(\Gamma\Gamma^*)^n = \{\lambda'_{mk}\}$ satisfies the column decay inequalities

$$|\lambda'_{mk}| \leqq C'\left(\frac{b}{a}\right)^{1-\delta} \kappa^{2n+1} m^{\delta-3/2} k^{1/2-\delta}. \tag{4.15}$$

We multiply again on the left, this time by Γ, and repeat the same argument; taking into account the symmetry of the matrix there results

$$\|(\Gamma\Gamma^*)^{n+1}\|_\delta \leqq C'\kappa^{2n+2}.$$

Thus if C' is suitably chosen, the first inequality in (4.13) follows by induction; the proof of the second is exactly the same.

LEMMA 4.4. *Let* $0<\delta\leqq\frac{1}{2}$. *Then*

$$\frac{4}{\pi}\int_0^\infty \frac{s^\delta\, ds}{(1+s^2)^2} < 1. \tag{4.16}$$

Proof. We have

$$\frac{4}{\pi}\int_0^\infty \frac{ds}{(1+s^2)^2} = 1, \qquad \frac{4}{\pi}\int_0^\infty \frac{s\, ds}{(1+s^2)^2} = \frac{2}{\pi},$$

and

$$\frac{4}{\pi}\int_0^\infty \frac{s^\delta\, ds}{(1+s^2)^2} \leqq \left(\frac{4}{\pi}\int_0^\infty \frac{s\, ds}{(1+s^2)^2}\right)^\delta \left(\frac{4}{\pi}\int_0^\infty \frac{ds}{(1+s^2)^2}\right)^{1-\delta}.$$

Combining Lemma 4.3 and 4.4 and using (4.2), we obtain the following.

PROPOSITION 4.1. *Let* $\Gamma=\{\gamma_{mk}\}$ *where* (4.11) *holds and* (i,j) $(i,j=0,1)$ *is fixed. Then the series* (4.3) *are convergent in the norm* (4.4) *for* $0<\delta\leqq\frac{1}{2}$.

The above discussion leads readily to the following.

THEOREM 4.1. *The solution to* (3.12) *is given by a system of the form*

$$\begin{aligned} A^j_{2n-1+i} &= \bar\alpha_{2n-1+i} + \sum_{k=1}^\infty \left(\lambda_{2n-1+i,2k-1+i}\bar\alpha_{2k-1+i} + \mu_{2n-1+i,2k-1+j}\bar\beta_{2k-1+j}\right), \\ B^i_{2n-1+j} &= \bar\beta_{2n-1+j} + \sum_{k=1}^\infty \left(\mu^*_{2n-1+j,2k-1+i}\bar\alpha_{2k-1+i} + \lambda^*_{2n-1+j,2k-1+j}\bar\beta_{2k-1+j}\right), \\ &\qquad\qquad i,j=0,1, \quad n=1,2,\cdots, \end{aligned} \tag{4.17}$$

and for $0<\delta\leqq\frac{1}{2}$,

$$\sup_{m,n}\left\{\left(|\lambda_{mn}|+|\mu_{mn}|+|\lambda^*_{mn}|+|\mu^*_{mn}|\right)\max\left(m^{\delta-1/2}n^{3/2-\delta}, m^{3/2-\delta}n^{\delta-1/2}\right)\right\} < \infty.$$

Remark. It is easily seen that if the infinite matrix $\Lambda = \{\lambda_{mn}\}$ has

$$\|\Lambda\|_\delta < \infty$$

for $0 < \delta \leqq \frac{1}{2}$ then

$$\sup_n \sum_{m=1}^{\infty} |m^\sigma \lambda_{mn} n^{-\sigma}|, \ \sup_m \sum_{n=1}^{\infty} |m^\sigma \lambda_{mn} n^{-\sigma}| < \infty$$

for $0 \leqq \sigma < \frac{1}{2}$. It follows then from Young's inequality that the matrix Λ induces an operator that is continuous with respect to the sequence space norm

$$|a| = \left(\sum_{n=1}^{\infty} n^{2\sigma} a_n^2 \right)^{1/2}$$

for $0 \leqq \sigma < \frac{1}{2}$. It follows therefore from Theorem 4.1 (cf. §3) that if $g \in H^k(\partial\Omega)$, $\frac{1}{2} \leqq k < 1$, then

$$\sum_{j=0}^{1} \sum_{n=1}^{\infty} n^{2k-1} \left(|A_n^j|^2 + |B_n^j|^2 \right) < \infty.$$

5. Truncation estimates. It is not difficult to see that the methods of the preceding section could be applied to obtain row and column decay estimates, uniform with respect to N, for the inverses of the matrices that appear in the truncated systems (3.14). Stronger results can be obtained by observing that, in view of (3.8) and (3.9), if $\gamma_{mk}^{(i,j)} = \gamma_{mk}$ is given by (4.11) and $\Gamma^{(i,j)} = \{\gamma_{mk}^{(i,j)}\}$ then the matrices $(-1)^{i+j}\Gamma^{(i,j)}$ have nonnegative entries. Thus, for example, for any of the matrices Γ, if Γ_N is a truncation of Γ then $\Gamma\Gamma^*$, $\Gamma_N\Gamma_N^*$ have nonnegative entries and the entries of $I + (\Gamma_N\Gamma_N^*) + \cdots + (\Gamma_N\Gamma_N^*)^n + \cdots$ do not exceed the corresponding entries of $I + (\Gamma\Gamma^*) + \cdots + (\Gamma\Gamma^*)^n + \cdots$. We conclude the following.

PROPOSITION 5.1. *Inversion of the finite system* (3.14) *yields a finite system of the form* (4.17) *with corresponding coefficients* $\lambda_{mn}(N)$, $\mu_{mn}(N)$, $\mu_{mn}^*(N)$, $\lambda_{mn}^*(N)$, $m, n = 1, \cdots, 2N$. *Moreover, if* $N' > N$ *then*

$$0 \leqq \lambda_{mn}(N) \leqq \lambda_{mn}(N') \leqq \lambda_{mn}, \ 0 \leqq \lambda_{mn}^*(N) \leqq \lambda_{mn}^*(N) \leqq \lambda_{mn}^*(N') \leqq \lambda_{mn}^*,$$
$$|\mu_{mn}(N)| \leqq |\mu_{mn}(N')| \leqq |\mu_{mn}|, \ |\mu_{mn}^*(N)| \leqq |\mu_{mn}^*(N')| \leqq |\mu_{mn}^*|.$$

Finally,

$$\sup_{m,n \leqq 2N} \left\{ |\lambda_{mn}(N)| + |\mu_{mn}(N)| + |\lambda_{mn}^*(N)| + |\mu_{mn}^*(N)| \right\} \times \max(m^{\delta - 1/2} n^{3/2 - \delta}, m^{3/2 - \delta} n^{\delta - 1/2}) \leqq C \tag{5.1}$$

for $0 < \delta \leqq \frac{1}{2}$, *where* $C = C(\delta)$ *is independent of* N.

Remark. Regarding the rate of convergence, it can be shown that

$$|\lambda_{mn}(N) - \lambda_{mn}|, \ |\mu_{mn}(N) - \lambda_{mn}|, \ |\mu_{mn}^*(N) - \mu_{mn}^*|, \ |\lambda_{mn}^*(N) - \lambda_{mn}^*| < C m^{1/2 - \delta} n^{1/2 - \delta} N^{2\delta - 2}, \qquad 1 \leqq m, n \leqq 2N, \tag{5.2}$$

for $0 \leqq \delta \leqq \frac{1}{2}$, where again C depends on δ but not on N.

In the absence of a better estimate than (5.2) for the rate of convergence of the inverses of the truncated matrices, in order to get better than $O(N^{-2})$ convergence of the solutions to (3.14) we must assume an a priori hypothesis concerning the decay of the coefficients A_n^j, B_n^j themselves. In the final sections we show when such an hypothesis can be verified.

In what follows, as in §3, $\bar{A}_n^j = \bar{A}_n^j(N)$, $\bar{B}_n^j = \bar{B}^j(N)$, $i,j=0,1$; $n=1,\cdots,2N$, denote the components of the solution to (3.14).

THEOREM 5.1. *Suppose that for some nonnegative number* σ, *the solution to* (3.12) *satisfies*

$$\sum_{i=0}^{1}\sum_{n=1}^{\infty} n^{2\sigma}\left(\left|A_n^i\right|^2+\left|B_n^i\right|^2\right)<\infty. \tag{5.3}$$

Then for $0<\delta\leqq\frac{1}{2}$,

$$\left|\bar{A}_n^i(N)-A_n^i\right|,\ \left|\bar{B}_n^i(N)-B_n^i\right|\leqq Cn^{1/2-\delta}N^{-1-\sigma+\delta}, \tag{5.4}$$

where $C=C(\delta)$.

Proof. Put

$$a_n^i(N)=\bar{A}_n^i(N)-A_n^i,\qquad b_n^i(N)=\bar{B}_n^i(N)-B_n^i.$$

The a_n^i, b_n^i satisfy the linear system with the same matrix that appears in (3.14) and the right-hand terms

$$\sum_{k=N+1}^{\infty} B_{2k-1+j}^i\left\langle\varphi_{2n-1+i}^j,\psi_{2k-1+j}^i\right\rangle=f_{2n-1+i}^j(N)$$

and

$$\sum_{k=N+1}^{\infty} A_{2k-1+i}^j\left\langle\psi_{2n-1+j}^i,\varphi_{2k-1+i}^j\right\rangle=h_{2n-1+j}^i(N),\qquad i,j=0,1,\quad n=1,\cdots,N.$$

From (3.9) and (5.3) we have, for example,

$$\begin{aligned}\left|f_n^j(N)\right|&\leqq Cn^{3/2}\sum_{k=N+1}^{\infty}k^{-5/2}B_{2k-1+j}^i,\\&\leqq Cn^{3/2}\left(\sum_{k=N+1}^{\infty}k^{-5-2\sigma}\right)^{1/2}\left(\sum_{k=1}^{\infty}k^{2\sigma}\left|B_{2k-1+j}^i\right|^2\right)^{1/2}\\&\leqq Cn^{3/2}N^{-2-\sigma},\end{aligned}$$

(here $(-1)^{n+i}=-1$). Similarly

$$\left|h_n^i(N)\right|\leqq Cn^{3/2}N^{-2-\sigma}.$$

We then apply Proposition 5.1, specifically inequality (5.1), to obtain (5.4).

6. Smooth solutions. We now show that if

$$g\in H^k(\partial\Omega),\qquad k\geqq\tfrac{3}{2}, \tag{6.1}$$

and if the solution u to (1.1), (1.2) belongs to $H^4(\Omega)$ or at least

$$\frac{\partial^2}{\partial y^2}\Delta u \in L^2(\Omega), \tag{6.2}$$

then the coefficients in the expansion (3.10) satisfy (5.3) with $\sigma = 2$.

Indeed if (6.2) holds, then by i) of Proposition A.1 in the Appendix we have

$$\frac{\partial^2}{\partial y^2}\Delta u = \sum_{i=0}^{1}\sum_{n=1}^{\infty}\left(A_n^i \Delta\varphi_n^i + B_n^i \Delta\psi_n^i\right)$$

where

$$\sum_{i=0}^{1}\sum_{n=1}^{\infty}\left(\left|A_n^i\right|^2 + \left|B_n^i\right|^2\right) < \infty.$$

Thus

$$u = v + w, \tag{6.3}$$

where

$$v = \frac{1}{\pi^2}\sum_{i=0}^{1}\sum_{n=1}^{\infty} n^{-2}\left(a^2 A_n^i \varphi_n^i - b^2 B_n^i \psi_n^i\right) \tag{6.4}$$

and

$$\Delta w = C_1 + C_2 x + C_3 y + C_4 xy. \tag{6.5}$$

To complete the proof of our assertion, we show that, given (6.1), we must have

$$C_1 = C_2 = C_3 = C_4 = 0. \tag{6.6}$$

LEMMA 6.1. *Suppose that u has the expansion* (3.10) *and* (5.3) *holds for some* σ *on the range*

$$0 \leqq \sigma < \tfrac{3}{2}. \tag{6.7}$$

Then

$$\frac{\partial u}{\partial \nu} \in H^{\sigma+1/2}(\partial\Omega). \tag{6.8}$$

LEMMA 6.2. *Suppose that*

$$w \in B_0(\Omega)$$

and (6.5) *holds. Then*

$$\frac{\partial w}{\partial \nu} \in H^k(\partial\Omega),$$

for $k \geqq \frac{3}{2}$ *implies* (6.6).

Proof of Lemma 6.1. From (3.9) it follows that the matrices $\Gamma = \{\gamma_{mn}\}$, Γ as in Lemma 4.3, satisfy

$$\sup_m m^\sigma \sum_{n=1}^{\infty} |\gamma_{mn}| n^{-\sigma}, \quad \sup_n n^{-\sigma} \sum_{m=1}^{\infty} |\gamma_{mn}| m^\sigma \leqq C$$

for $0 \leqq \sigma < \frac{3}{2}$. It then follows from (5.3) and Young's inequality that

$$\sum_{i=0}^{1} \sum_{n=1}^{\infty} n^{2\sigma}\left(|\overline{\alpha}_{n,i}|^2 + |\overline{\beta}_{n,i}|^2\right) < \infty,$$

provided (6.7) holds. Thus (cf. §3) if (6.7) holds then (5.3) implies (6.8).

Proof of Lemma 6.2. If we decompose u as

$$u = \sum_{i=0}^{1} \sum_{j=0}^{1} u^{(i,j)}, \tag{6.9}$$

where

$$u^{(i,j)}(x,y) = \tfrac{1}{4}\big[u(x,y) + (-1)^i u(a-x,y) + (-1)^j u(x,b-y) + (-1)^{i+j} u(a-x,b-y)\big], \tag{6.10}$$

then it is clear that the smoothness properties (6.1) ($g = \partial u/\partial \nu$) and (6.2) are invariant under $u \mapsto u^{(i,j)}$, $i,j = 0,1$. Let $w^{(i,j)} \in B_0(\Omega)$ be determined by

$$\Delta w^{(i,j)}(x,y) = \left(\frac{2}{a}x - 1\right)^i \left(\frac{2}{b}y - 1\right)^j.$$

Then

$$w^{(i,j)}(x,y) = \sum_{n=1}^{\infty} \tilde{A}^j_{2n-1+i} \varphi^j_{2n-1+i} + \sum_{n=1}^{\infty} \tilde{B}^i_{2n-1+j} \psi^i_{2n-1+j}, \tag{6.11}$$

where

$$\tilde{A}^j_n = (-1)^{n+j} \frac{a}{c^j_n\big[(-1)^j \cosh(n\pi b/a) + 1\big]} \left(\frac{2}{\pi n}\right)^{3/2}, \tag{6.12}$$

$$\tilde{B}^j_n = (-1)^{n+j} \frac{b}{d^j_n\big[(-1)^j \cosh(n\pi a/b) + 1\big]} \left(\frac{2}{\pi n}\right)^{3/2}, \tag{6.13}$$

cf. (3.5), (3.6). We find the coefficients α_n, β_n in the expansion (2.5) of $g = (\partial w^{(i,j)}/\partial \nu)$ as follows. The coefficients $\overline{A}^j_n$, $\overline{B}^j_n$ in (6.11) satisfy equations (3.12) with the $\overline{\alpha}_{n,j}$, $\overline{\beta}_{n,j}$ on the right defined in terms of the α_n, β_n as indicated in §3. A check of the signs in (3.9) and (6.12), (6.13) shows that in all of the resulting equations (3.12) the first term on the left and the summation always agree in sign. Thus, for example

$$|\tilde{A}^j_{2n-1+i}| \leqq \left|\tilde{A}^j_{2n-1+i} + \sum_{k=1}^{\infty} \tilde{B}_{2k-1+j} \left\langle \varphi^j_{2n-1+i}, \psi^i_{2k-1+j} \right\rangle\right|.$$

It readily follows from (6.12), (6.13), that

$$\frac{\partial w^{(i,j)}}{\partial \nu} \notin H^k(\partial\Omega)$$

for $k \geqq \frac{3}{2}$.

To complete the proof, we consider one of the $u^{(i,j)}$ given by (6.10). If we make the decomposition (6.3), as indicated above, of $u^{(i,j)}$, then the second term w is of the form $Cw^{(i,j)}$; however $(\partial u^{(i,j)}/\partial \nu) \in H^k(\partial\Omega)$, for some $k \geqq \frac{3}{2}$; by (6.4) and Lemma 6.1, the first term v has $(\partial v/\partial \nu) \in H^2(\partial\Omega)$ and thus we must have $C=0$.

We summarize the above observations in the following.

PROPOSITION 6.1. *Let* (6.1) *hold and suppose that the solution u to* (1.1), (1.2) *satisfies* (6.2). *Then the coefficients in the expansion* (3.10) *of u satisfy* (5.3) *with*

$$\sigma = 2.$$

7. The clamped plate. The problem of the clamped plate is

$$\Delta^2 u = f \quad \text{in } \Omega, \tag{7.1}$$

$$u = \frac{\partial u}{\partial \nu} = 0 \quad \text{on } \partial\Omega. \tag{7.2}$$

The formula of Zaremba, in its usual formulation, (cf. e.g. [1], a more general formulation is given in [4]) expresses the Green's function Γ for (7.1), (7.2) as

$$\Gamma = \Gamma_1 - \Gamma_2, \tag{7.3}$$

where Γ_1 is the Green's function for the differential equation (7.1) and boundary conditions

$$u = \Delta u = 0 \quad \text{on } \partial\Omega; \tag{7.4}$$

(in the case of a rectangle, or more generally a polygon, conditions (7.4) are the boundary conditions of the simply supported plate) and Γ_2 is the reproducing kernel of the space $B_0(\Omega)$. Thus the solution u to (7.1), (7.2) decomposes as

$$u = u_1 - u_2 \tag{7.5}$$

where u_1 is obtained by solving (7.1), (7.4) and u_2 can be computed using Γ_2 when the latter is known. Alternatively, u_2 can be found by solving (1.1), (12) with

$$g = \frac{\partial u_1}{\partial \nu}.$$

Note that the representation of u_1 as a double sine series can be written immediately once a similar representation for f is obtained.

It is easily seen that if

$$f \in L^2(\Omega), \tag{7.6}$$

then the solution u_1 to (7.1), (7.4) belongs to $H^4(\Omega)$. Moreover, an argument similar to that indicated for the proof of assertion (*) in §2 shows that (7.6) implies

$$\frac{\partial u_1}{\partial \nu} \in H^{5/2}(\partial\Omega). \tag{7.7}$$

Finally, in the case of a rectangle Ω the solution to (7.1), (7.2) is known to belong to $H^4(\Omega)$ when (7.6) holds [7]. We conclude that $u_2 \in H^4(\Omega)$ and thus in view of (7.7) u_2 satisfies the hypothesis of Proposition 6.1. If u_2 is to be found by the method described above, then the truncation estimates (5.4) hold with $\sigma = 2$.

8. Example. We consider now the special case of the problem (7.1), (7.2) in which Ω is the square of side π and

$$f \equiv 1 \quad \text{in } \Omega.$$

This computation has been performed repeatedly before, [5], [12], [13]. We are primarily interested here in demonstrating how the coefficients of the truncated reduced problem vary with N.

Following Hencky [5] (see also [9]), we write the solution u as

$$u = w - v$$

where

$$w(x,y) = \tfrac{1}{8}xy(\pi - x)(\pi - y) \tag{8.1}$$

and v is biharmonic on Ω with

$$v = 0, \qquad \frac{\partial v}{\partial \nu} = \frac{\partial w}{\partial \nu} \quad \text{on } \partial\Omega.$$

Because of the symmetries involved we see that v can be represented (with the notation of §3 and with $a = b = \pi$) in the form

$$v(x,y) = \sum_{k=1}^{\infty} A_{2k+1}\left(\varphi^0_{2k+1}(x,y) + \psi^0_{2k+1}(x,y)\right).$$

The equations that determine the A_n are

$$A_n = \sqrt{\frac{1+\cosh n\pi}{n\pi + \sinh n\pi}}\left[\frac{\sqrt{2\pi}}{n^{5/2}} - \frac{8n^{3/2}}{\pi}\sum \frac{m^{3/2}}{(m^2+n^2)^2}\sqrt{\frac{1+\cosh m\pi}{m\pi + \sinh m\pi}}\, A_m\right], \tag{8.2}$$

$n = 1, 3, 5, \cdots$; the summation on the right in (8.2) is over the positive odd integers.

The form in which (8.2) has been written lends itself to an obvious iterative procedure (a minor variant of Gauss–Seidel) for solution of the truncated systems. Because the matrix need not be stored, this can even be carried out on a programmable calculator such as e.g. the Texas Instrument TI-59. The number of computations can be reduced by making the change of variable

$$\bar{A}_n = \sqrt{\frac{1+\cosh n\pi}{n\pi + \sinh n\pi}}\, A_n$$

in (8.2).

Using the TI-59 the truncated systems involving the first 8, 15, 20, 25 equations were solved (it would be possible, but because of the time involved not practical, to handle up to 50 equations). These systems were solved consecutively and, for example, the first solution (of the 8×8) was augmented with 7 zeros and used as the initial point for the iterative solution of the 15×15 system. The values obtained for the A_n are

shown in Table 1.

TABLE 1

n	$N=8$	$N=15$	$N=20$	$N=25$
1	1.503134712	1.503134586	1.503134725	1.503134725
3	−.0178766112	−.0178752812	−.0178753017	−.0178753021
5	−.0096545055	−.0096503228	−.0096503056	−.0096503093
7	−.0040740937	−.0040665724	−.0040665248	−.0040665347
9	−.0019191169	−.0019086107	−.0019085294	−.0019085479
11	−.0009972989	−.0009845646	−.0009844483	−.0009844771
13	−.0005594979	−.0005453758	−.0005452258	−.0005452658
15	−.0003333588	−.0003185945	−.0003184141	−.0003184658
17		−.0001936597	−.0001934535	−.0001935167
19		−.0001212241	−.0001209974	−.0001210715
21		−.0000774764	−.0000772344	−.0000773185
23		−.0000501694	−.0000499171	−.0000500102
25		−.000032663	−.0000324049	−.0000325056
27		−.0000211942	−.0000209342	−.0000210414
29		−.0000135498	−.000013291	−.0000134034
31			−.0000081303	−.0000082467
33			−.0000046124	−.0000047318
35			−.0000022004	−.0000023216
37			−.0000005436	−.0000006657
39			.0000005913	.000000469
41				.0000012401
43				.0000017554
45				.0000020898
47				.0000022958
49				.0000024105

Appendix. Here we shall fill in briefly the theoretical details on which the preceding development was based.

Let $H(\Omega)$ denote the subspace of $L^2(\Omega)$ that consists of harmonic functions and let $H_1(\Omega)$ and $H_2(\Omega)$ denote respectively the subspaces of $H(\Omega)$ that are spanned by the sets

$$S_1 = \{\Delta\varphi_n^i(x,y): i=0,1; n=1,2,\cdots\},$$

and

$$S_2 = \{\Delta\psi_n^i(x,y): i=0,; n=1,2,\cdots\};$$

cf. (3.5), (3.6).

PROPOSITION A.1. *The following are equivalent*:

i) $H(\Omega)$ *admits the direct sum representation*

$$H(\Omega) = H_1(\Omega) + H_2(\Omega);$$

ii) *every* $u \in B_0(\Omega)$ *is uniquely representable in the form* (3.10) *where* (3.11) *holds*;

iii) *for each pair* (i,j) *the matrix* $\Gamma = \{\langle \varphi_{2m-j}^i, \psi_{2h-1+i}^{1-j} \rangle\}$ *satisfies*

$$\|\Gamma\| < 1.$$

Proof. Recall that the sets S_1 and S_2 are orthonormal in $L^2(\Omega)$. The equivalence of i) and ii) is immediate from the representation (2.2). To see the equivalence of i) and iii), one notes that the matrices Γ can be used to represent the restriction to $H_1(\Omega)$ of the orthogonal projection onto $H_2(\Omega)$ with respect to the orthonormal bases S_1 and S_2.

Remark. Assertion i) in Proposition A.1 is proved in [2].

PROPOSITION A.2. *The following are equivalent*:

i) *every* $u \in B_0(\Omega)$ *is uniquely representable in the form* (3.10) *where* (3.11) *holds*;

ii) *if* $u \in B_0(\Omega)$ *then there exists a* $v \in B_0$ *such that*

$$\frac{\partial v}{\partial \nu} = \frac{\partial u}{\partial \nu} \quad \text{on } \{(x,y): 0<x<a, y=0 \text{ or } y=b\},$$

and

$$\frac{\partial v}{\partial \nu} = 0 \quad \text{on } \{(x,y): 0<y<b, x=0 \text{ or } x=a\};$$

iii) *the problem* (1.1), (1.2) *has a solution* $u \in B_0(\Omega)$ *if and only if*

$$g \in H^{1/2}(\partial\Omega). \tag{A.1}$$

Proof. We have already noted ((*) §2) the necessity of (A.1). To see that i) implies ii), let i) be assumed; then if $u \in B_0(\Omega)$ has the representation (3.10),

$$u \mapsto \sum_{j=0}^{1} \sum_{n=1}^{\infty} A_n^j \psi_n^j$$

is a bounded projection in $B_0(\Omega)$, call it P. Using Green's theorem, one easily sees that for $u \in B_0(\Omega)$, $v = P^* u$ satisfies the conditions of ii); (P^* denotes the adjoint of P). Conversely, if ii) holds, then it follows from the necessity of (A.1) and the closed graph theorem that $u \mapsto v$ is a bounded projection. The converse implication follows readily, cf. [4]. The implication iii) implies ii) is immediate. Conversely, suppose data (2.5), satisfying (2.6), are given. For simplicity assume $\beta_n(0) = \beta_n(a) = 0$, $n = 1, 2, \cdots$. One can then obviously find w of the form

$$w = \sum_{i=0}^{1} \sum_{n=1}^{\infty} A_n^i \varphi_n^i, \qquad \sum_{i=0}^{1} \sum_{n=1}^{\infty} A_n^2 < \infty,$$

such that

$$\frac{\partial w}{\partial \nu} = g \quad \text{on } \{(x,y): 0<x<a, y=0 \text{ or } y=b\},$$

the assumption ii) then implies the existence of a $u \in B_0(\Omega)$ whose normal derivative coincides with the given data.

Remark. As we noted in [4], the assertion ii) of Proposition A.2 follows readily from the existence theorem for the first biharmonic boundary value problem on a rectangle that is stated in [2]; the proof of that theorem was given in [3].

We shall now sketch an elementary proof of assertion ii) of Proposition A.2.

LEMMA A.1. *Let* $v(x,y)$ *be of class* C^∞ *on a neighborhood of*

$$s = \{(x,y): 0 \leqq x, y, x^2 + y^2 < \varepsilon^2\}$$

and such that

i) $$v(0,y)=0, \quad 0\leqq y\leqq \varepsilon, \quad v(x,0)=0, \quad 0\leqq x\leqq \varepsilon,$$

and

ii) $$v(x,y)=0, \qquad x^2+y^2\geqq \tfrac{1}{2}\varepsilon^2.$$

Let

$$w(x,y)=\tan^{-1}\frac{y}{x}v(x,y), \qquad x,y\in s.$$

Then

$$\iint_s |\Delta w(x,y)|^2\,dx\,dy\leqq C\iint_s |\Delta v(x,y)|^2\,dx\,dy,$$

where C is an absolute constant.

Proof. We have

$$\text{(A.2)} \qquad \iint_s |\Delta v|^2\,dx\,dy\Gamma\int_0^{\pi/2}\int_0^{\varepsilon}\left[\frac{1}{r}\frac{\partial}{\partial r}r\frac{\partial v}{\partial r}+\frac{1}{r^2}\frac{\partial^2 v}{\partial\theta^2}\right]^2 r\,dr\,d\theta.$$

Since v is of class C^∞ on a neighborhood of $(0,0)$, it follows from i) that

$$\lim_{r\to 0}\frac{1}{r}\frac{\partial^2 v}{\partial\theta\,\partial r}\frac{\partial v}{\partial\theta}=0.$$

Using this and integrating by parts, we get

$$\text{(A.3)} \quad \int_0^{\pi/2}\int_0^{\varepsilon}\left(\frac{\partial}{\partial r}r\frac{\partial v}{\partial r}\right)\frac{1}{r^2}\frac{\partial^2 y}{\partial\theta^2}\,dr\,d\theta$$
$$=\int_0^{\pi/2}\int_0^{\varepsilon}\left[\frac{1}{r}\left(\frac{\partial^2 v}{\partial r\,d\theta}\right)^2-\frac{2}{r^2}\frac{\partial v}{\partial\theta}\frac{\partial^2 v}{\partial\theta\,\partial r}\right]dr\,d\theta.$$

We have

$$\text{(A.4)} \quad \int_0^{\pi/2}\int_0^{\varepsilon}\frac{2}{r^2}\frac{\partial v}{\partial\theta}\frac{\partial^2 v}{\partial\theta\,\partial r}\,dr\,d\theta\leqq\int_0^{\pi/2}\int_0^{\varepsilon}\left[\frac{1}{r}\left(\frac{\partial^2 v}{\partial r\,\partial\theta}\right)^2+\frac{1}{r^3}\left(\frac{\partial v}{\partial\theta}\right)^2\right]dr\,d\theta$$
$$\leqq\int_0^{\pi/2}\int_0^{\varepsilon}\left[\frac{1}{r}\left(\frac{\partial^2 v}{\partial r\,d\theta}\right)^2+\frac{1}{16r^3}\left(\frac{\partial^2 v}{\partial\theta^2}\right)^2\right]dr\,d\theta;$$

the last step follows by noting that $\int_0^{\pi/2}(\partial v/\partial\theta)\,d\theta=0$ for any r and expanding $\partial^2 v/\partial\theta^2$ in a Fourier series. From (A.3) and (A.4) we have

$$\text{(A.5)} \qquad -\frac{1}{8}\int_0^{\pi/2}\int_0^{\varepsilon}\frac{1}{r^3}\left(\frac{\partial^2 v}{\partial\theta^2}\right)^2 dr\,d\theta\leqq 2\int_0^{\pi/2}\int_0^{\varepsilon}\left(\frac{\partial}{\partial r}r\frac{\partial v}{\partial\theta}\right)\frac{1}{r^2}\frac{\partial^2 v}{\partial\theta^2}\,dr\,d\theta.$$

Upon expanding the right-hand side of (A.2) and using (A.5) we get

$$\text{(A.6)} \qquad \int_0^{\pi/2}\int_0^{\varepsilon}\frac{1}{r^3}\left(\frac{\partial^2 v}{\partial\theta^2}\right)^2 dr\,d\theta\leqq 2\int_0^{\pi/2}\int_0^{\varepsilon}|\Delta v|^2 r\,dr\,d\theta.$$

Now

$$\Delta w = \Delta(\theta v) = \theta \Delta v + \frac{2}{r^2}\frac{\partial v}{\partial \theta},$$

so that the asserted inequality follows from (A.6) and the inequality

$$\int_0^{\pi/2}\int_0^{\varepsilon} \frac{1}{r^3}\left(\frac{\partial v}{\partial \theta}\right)^2 dr\,d\theta \leqq \frac{1}{16}\int_0^{\pi/2}\int_0^{\varepsilon} \frac{1}{r^3}\left(\frac{\partial^2 v}{\partial \theta^2}\right) dr\,d\theta.$$

Let $X(\Omega)$ denote the set of functions of the form

$$u(x,y) = \sum_{m=1}^{\infty}\sum_{n=1}^{\infty} a_{mn} \sin\frac{m\pi x}{a}\sin\frac{n\pi y}{b} \tag{A.7}$$

and let

$$\|u\|^2 = \frac{\pi^2}{4ab}\sum_{m=1}^{\infty}\sum_{n=1}^{\infty}(a^2n^2+b^2m^2)^2 a_{mn}^2 < \infty,$$

(note that $\|u\|^2 = \int\int_{\Omega} |\Delta u|^2\,dx\,dy$). With the norm $\|\cdots\|$, $X(\Omega)$ is a Hilbert space and $B_0(\Omega)$ is a subspace. The orthogonal complement of $B_0(\Omega)$ in $X(\Omega)$ is the space $X_0(\Omega)$ of the clamped plate (i.e., the completion in $X(\Omega)$ of $C_0^{\infty}(\Omega)$)

$$X(\Omega) = B_0(\Omega) \oplus X_0(\Omega),$$

and $u \in X_0(\Omega)$ implies

$$u = \frac{\partial u}{\partial \nu} = 0 \quad \text{on } \partial\Omega.$$

Let Q denote the orthogonal projection of $X(\Omega)$ onto $B_0(\Omega)$. Then if $w \in X(\Omega)$ and $\partial w/\partial \nu = g$ on $\partial\Omega$ then $u = Qw$ is the unique solution to (1.1), (1.2). We claim that it is possible to construct an operator A on $X(\Omega)$ such that if $u = Aw$ then $\partial w/\partial \nu$ and $\partial u/\partial \nu$ agree on the horizontal segments of $\partial\Omega$ while $\partial u/\partial \nu = 0$ on the vertical segments. The operator P^* described above is then obtained as the restriction to $B_0(\Omega)$.

Let $w \in X(\Omega)$ and be the sum of a finite series of the form (A.7). Using a partition of the identity, we can write

$$w = u + v$$

where u satisfies the hypothesis of Lemma A.1 and

$$\|u\|,\ \|v\| \leqq C\|w\|,$$

where C depends only on the partition. From Lemma A.1 it follows that

$$w_1 = \left(1 - \frac{2}{\pi}\tan^{-1}\frac{y}{x}\right) w \in X_0(\Omega)$$

and

$$\|w_1\| \leqq C'\|w\|.$$

Note that

$$\frac{\partial w_1}{\partial \nu} = 0$$

on the left-hand vertical segment of $\partial\Omega$ while $\partial w_1/\partial\nu$ nd $\partial u/\partial\nu$ agree on the lower horizontal segment. The construction of the specified operator A can now be carried out using appropriate partitions of identity and the above construction at the corners. We thus can prove the following.

PROPOSITION A.3. *There exists a bounded projection P^* on $B_0(\Omega)$ such that for any $u \in B_0(\Omega)$ if $v = P^*u$ then $\partial v/\partial\nu$ and $\partial u/\partial\nu$ agree on the horizontal segments of $\partial\Omega$ while*

$$\frac{\partial v}{\partial \nu} = 0$$

on the vertical segments. Thus each of the equivalent assertions of Propositions A.1 *and* A.2 *is valid.*

REFERENCES

[1] N. ARONSZAJN, *Theory of reproducing kernels*, Trans. Amer. Math. Soc., 68 (1950), pp. 337–404.

[2] N. ARONSZAJN, R. D. BROWN AND R. S. BUTCHER, *Construction of the solution of boundary value problems for the biharmonic operator in a rectangle*, Ann. Inst. Fourier, 23 (1973), pp. 49–89.

[3] N. ARONSZAJN AND P. SZEPTYCKI, *Theory of Bessel potentials. Part* IV. *Potentials of subcartesian spaces with singularities of polyhedral type*, Ann. Inst. Fourier, 25 (1975), pp. 27–69.

[4] C. V. COFFMAN, *The reproducing kernel space of the clamped plate*, Contributions to Analysis and Geometry, D. N. Clark, G. Pecelli and R. Sacksteder, eds., Johns Hopkins Univ. Press, Baltimore, 1981.

[5] H. HENCKY, *Der Spannungeszustand in rechteckigen Platten*, R. Oldenbourg, Munich, 1913.

[6] S. H. GOULD, *Variational Methods for Eigenvalue Problems: An Introduction to the Weinstein Method of Intermediate Problems*, Univ. Toronto Press, Toronto, 1966.

[7] V. A. KOLDORKINA, *On the solutions of the equation $\Delta\Delta u = f$ in a piecewise smooth plane domain*, Differencial'nye Uravnenija, 8 (1972), pp. 374–376; Differential Equations, 8 (1972), pp. 285–287.

[8] G. LARUICELLA, *Sur l'intégration de l'équation à l'équilibre des plaques élastiques encastrées*, Acta Math., 32 (1909), pp. 201–256.

[9] A. E. H. LOVE, *A Treatise on the Mathematical Theory of Elasticity*, Dover, New York, 1944.

[10] S. G. MIKHLIN, *Variational Methods in Mathematical Physics*, Pergamon, Oxford, 1964.

[11] E. H. RAFAL'SON, *A problem arising in the solution of the biharnomic equation*, Dokl. Akad. Nauk SSSR, 64 (1949), pp. 799–802. (In Russian.)

[12] S. TIMOSHENKO, *Theory of Plates and Shells*, McGraw-Hill, New York, 1940.

[13] _____, *Bending of rectangular plates with clamped edges*, Proc. 5th International Congress for Applied Mechanics, Cambridge, MA, John Wiley, New York, 1939.

SIAM J. MATH. ANAL.
Vol. 17, No. 2, March 1986

008

ABSTRACT NONLINEAR VOLTERRA EQUATIONS WITH POSITIVE KERNELS*

NORIMICHI HIRANO†

Abstract. The existence of the solutions of the equation

$$u(t)+\int_0^t a(t-s)Au(s)\,ds \ni f(t), \qquad 0 \leqq t \leqq T,$$

in a Hilbert space, is proved, where $A : H \to H$ is a pseudo-monotone operator and $f \in L^2(0, T;\ H)$.

1. Introduction. In this paper we consider the existence of solutions to the abstract Volterra integral equation

$$u(t)+\int_0^t a(t-s)Au(s)\,ds \ni f(t), \qquad 0 \leqq t \leqq T, \tag{1.1}$$

in a Hilbert space H, where A is a nonlinear (possibly multivalued) operator from H into itself, $a(t)$ is a real function on $[0, T]$, f is a function from $[0, T]$ into H. General existence results for (1.1) were obtained by Barbu [1], Crandall and Nohel [5], and Gripenberg [6]. Recently Kiffe and Stecher [7], [8] obtained existence results for (1.1) in case A is maximal monotone without assuming any continuity condition on A or differentiability conditions on f. In the present paper, we also consider the problem (1.1) without assuming any continuity on A or differentiability condition on f. Our purpose in this paper is to establish existence results for the problem (1.1) under the assumption that A is a pseudo-monotone operator on H. It is known that a broad class of (multivalued) operators satisfy the pseudo-monotonicity. To prove our results we make use of an existence result for pseudo-monotone operators [3].

In §2, we give notation, definitions, and the statements of our results. In §3, the proofs of the results are given. We give examples in §4.

2. Statement of results. Throughout this paper H will denote a real Hilbert space with the norm denoted $|\cdot|$ and the inner product $(\cdot, \cdot)$. We will denote by $L^2(0, T;\ H)$ the space of all H-valued function $u : [0, T] \to H$ such that $\int_0^T |u(t)|^2\,dt < +\infty$ and by $L^\infty(0, T;\ H)$ the space of H-valued function such that $\operatorname{ess\,sup}_{t \in [0, T]} |f(t)| < \infty$. The norm and inner product of $L^2(0, T;\ H)$ will be denoted by $\|\cdot\|$ and $\langle \cdot, \cdot \rangle$. Strong and weak convergence in $L^2(0, T;\ H)$ is denoted by "$\to$" and "$\rightharpoonup$" respectively. Let A be a nonlinear (multivalued) operator from H into itself. We shall denote by $D(A)$ the domain of A, i.e., $D(A) = \{x \in H;\ Ax \neq \emptyset\}$. For each $z \in H$, A_z denotes the operator defined by $A_z x = A(x+z)$, for $x \in D(A)$. The multivalued mapping $A : H \to H$ is said to be pseudo-monotone on H [3], if A satisfies the following conditions:

(1) $D(A) = H$.

(2) For any sequence $\{u_n\}$ and $\{w_n\}$ in H such that u_n converges weakly to u, $w_n \in Au_n$ for each $n \geqq 1$, and $\limsup_{n \to \infty} (w_n, u_n - u) \leqq 0$, and for any $v \in H$, there exists $w \in Au$ such that

$$(w, u-v) \leqq \liminf_{n \to \infty} (w_n, u_n - v). \tag{2.1}$$

*Received by the editors August 18, 1983, and in revised form January 25, 1984.

† Department of Mathematics, Yokohama National University, 156, Tokiwadai, Hodogaya, Yokohama, Japan.

The mapping $A: H \to H$ is said to be finitely continuous if Ax is closed convex subset of H for each $x \in H$ and for each finite set G of H, A is upper semicontinuous mapping from the closed convex hull of G into 2^H, with respect to the weak topology of H. The mapping A is said to be monotone if

$$(x_1 - x_2, y_1 - y_2) \geqq 0$$

for $x_i \in H$, $y_i \in Ax_i$ $(i = 1, 2)$. A monotone mapping A is said to be maximal monotone if the range of $I + \lambda A$ is H for all $\lambda > 0$. It is known that a maximal monotone operator with $D(A) = H$ is finitely continuous and pseudo-monotone (cf. [3]). A mapping $A: H \to H$ is said to be completely continuous if it is continuous from the weak topology of H to the strong topology of H. For each $a \in L^1(0, T)$, L_a denotes a linear continuous operator from $L^2(0, T)$ into itself defined by

$$(L_a f)(t) = \int_0^t a(t-s) f(s)\, ds \quad \text{for } 0 \leqq t \leqq T$$

for each $f \in L^2(0, T)$. Then the adjoint operator L_a^* is given by

$$(L_a^* f)(t) = \int_t^T a(s-t) f(s)\, ds \quad \text{for } 0 \leqq t \leqq T.$$

We now state the assumptions for the kernel function $a(t)$.

(2.2) $a(t) \in L^1(0, T)$ is of positive type on $[0, T]$, i.e., for each $f \in L^2(0, T)$,

$$\int_0^t f(\tau) \int_0^\tau a(\tau - s) f(s)\, ds\, d\tau \geqq 0 \quad \text{for } 0 \leqq t \leqq T.$$

(2.3) L_a^* is injective, i.e., $L_a^* f = 0$ means $f = 0$.

Sufficient conditions for $a(t)$ to satisfy (2.2) are given in [10] and [11]. It is easy to see that $a(t)$ satisfies (2.3) if $a(0) \neq 0$, $a \in C(0, T)$ and $a' \in L^1(0, T)$. It is also easy to see that the sum of a function of positive type $b(t)$ and a positive constant c, $a(t) = b(t) + c$ satisfies (2.3).

Let $A: H \to H$ be a multivalued operator with $D(A) = H$. Denote by $\tilde{A}$ the operator defined by

$$\tilde{A}u = \{ w \in L^2(0, T;\ H) : w(t) \in Au(t) \text{ a.e. on } [0, T] \},$$

for each $u \in L^2(0, T;\ H)$. It is well known that if A is a maximal monotone operator on H, then $\tilde{A}$ is a maximal monotone operator on $L^2(0, T;\ H)$. We now state basic assumptions for the nonlinear operator A on H.

(2.4) There exist constants c_1, c_2 such that

$$|y| \leqq c_1 + c_2 |x| \quad \text{for all } y \in Ax,\ x \in H,$$

(2.5) $$\liminf_{y \in Ax,\ |x| \to \infty} (x, y)/|x| > -\infty.$$

Remark. The condition (2.5) guarantees that for each $\varepsilon > 0$, $\varepsilon I + A$ is coercive, i.e., $\lim_{y \in \varepsilon x + Ax,\ |x| \to \infty} \langle x, y \rangle / |x| = +\infty$. We note that if A is monotone, then A satisfies (2.5).

We next state assumptions for the nonlinear operator A on $L^2(0,T;\ H)$ corresponding to (2.4) and (2.5).

$(2.4)'$ There exist constants c_1', c_2' such that

$$\|v\| \leqq c_1' + c_2' \|u\| \quad \text{for all } v \in Au,\ u \in L^2(0,T;\ H),$$

$(2.5)'$ $$\liminf_{v \in Au,\ \|u\| \to \infty} (v,u)/\|u\| > -\infty.$$

Remark. If a maximal monotone operator $A : H \to H$ with $D(A) = H$ satisfies (2.4), then we can see that $D(\tilde{A}) = L^2(0,T;\ H)$. It is obvious that if an operator $A : H \to H$ satisfies (2.4), then $\tilde{A}$ satisfies $(2.4)'$. It is also easy to see that $\tilde{A}$ satisfies $(2.5)'$ whenever A satisfies (2.4) and (2.5).

THEOREM 1. *Let A be a finitely continuous pseudo-monotone operator on $L^2(0,T;\ H)$ and satisfy $(2.4)'$, $(2.5)'$. Suppose that $a(t)$ satisfies (2.2) and $c_2' \|a\|_{L^1(0,T)} < 1$. Then for each $f \in L^2(0,T;\ H)$, the equation*

$$u(t) + \int_0^t a(t-s)(Au)(s)\, ds \ni f(t), \qquad 0 \leqq t \leqq T, \tag{2.6}$$

has a solution on $[0,T]$.

Remark. It is known that the sum of monotone operator and completely continuous operator, $A = B + C$ is pseudo-monotone. Then Theorem 1 is applicable to the existence problems of local solutions of equations of the form

$$\frac{du}{dt} + A(t)u(t) + \int_0^T k(t,s) Bu(s)\, ds \ni 0, \qquad u(0) = u_0, \tag{2.7}$$

where $A(t) : H \to H$ is a maximal monotone operator for each $0 \leqq t \leqq T$, $B : H \to H$ is completely continuous operator, $u_0 \in H$ and $k(t,s) \in L^2([0,T] \times [0,T])$. In fact, by integrating the above differential equation, the problem (2.7) arrives at the problem (2.6).

COROLLARY 1. *Let $A : H \to H$ be a multivalued operator satisfying (2.4) and (2.5). Suppose that for each $t \in [0,T]$, $\tilde{A}$ is finitely cotinuous and pseudo-monotone on $L^2(0,t;\ H)$. Then for each $f \in L^2(0,T;\ H)$, (1.1) has a solution on $[0,T]$.*

COROLLARY 1′ (Kiffe and Stecher [7]). *Let $A : H \to H$ be a maximal monotone operator with $D(A) = H$ and satisfy (2.4). Let $a(t)$ satisfy (2.2). Then for each $f \in L^2(0,T;\ H)$, (1.1) has a solution on $[0,T]$.*

Remark. Corollary 1 is a reformulation of Theorem 1 in case where A is an operator on H and we can obtain Kiffe and Stecher's result (Corollary 1′) from Corollary 1. But we can see that Theorem 1 (or Corollary 1) is not very effective for operators on H because, except for monotone operators, operators on H rarely satisfy the assumption that $\tilde{A}$ is pseudo-monotone on $L^2(0,t;\ H)$. Thus it is desirable to establish an existence result under the assumption that $A : H \to H$ is pseudo-monotone on H. We next state our main result.

THEOREM 2. *Let $A : H \to H$ be a finitely continuous pseudo-monotone operator on H satisfying (2.4), (2.5). Suppose that $a(t) \in L^2(0,T)$ satisfies (2.2) and (2.3). Then for each $f \in L^2(0,T;\ H)$, (1.1) has a solution on $[0,T]$.*

COROLLARY 2. *Let $A : H \to H$ be a maximal monotone operator with $D(A) = H$ and $B : H \to H$ be a completely continuous operator with $D(B) = H$. Suppose that $A + B$ satisfy (2.4) and (2.5) with A replaced by $A + B$. Let $a(t) \in L^2(0,T)$ satisfy (2.2) and (2.3).*

Then for each $f \in L^2(0,T;H)$, *the equation*

$$u(t)+\int_0^t a(t-s)(Au(s)+Bu(s))\,ds \ni f(t), \qquad 0 \leqq t \leqq T, \tag{2.8}$$

has a solution on $[0,T]$.

3. Proofs. We first give fundamental lemmas. The following lemma is a direct consequence of the definition of pseudo-monotonicity.

LEMMA 1. *Let* F *be a real Hilbert space and* A *be a pseudo-monotone operator on* F. *Let* $\{u_n\}$ *and* $\{w_n\}$ *be sequences such that* u_n *converges weakly to* u, $w_n \in Au_n$ *for each* $n \geqq 1$ *and* $\liminf_{n\to\infty}(w_n, u_n-u) \leqq 0$. *Then* $\liminf_{n\to\infty}(w_n, u_n-u)=0$.

Proof. Choose a subsequence $\{u_{n_i}\}$ of $\{u_n\}$ such that

$$\lim_{i\to\infty}(w_{n_i}, u_{n_i}-u)=\liminf_{n\to\infty}(w_n, u_n-u).$$

Then from the definition of pseudo-monotonicity, there exists $w \in Au$ such that

$$0=(w,u-u) \leqq \lim_{i\to\infty}(w_{n_i}, u_{n_i}-u).$$

Thus we obtain that $\liminf_{n\to\infty}(w_n, u_n-u)=0$.

LEMMA 2. *Let* F *be a real Hilbert space and* A *be a finitely continuous pseudo-monotone operator on* F. *Let* $\{u_n\}$ *and* $\{w_n\}$ *be sequences such that* u_n *converges weakly to* u, $w_n \in Au_n$ *for each* $n \geqq 1$ *and* $\limsup_{n\to\infty}(w_n, u_n-u) \leqq 0$. *Suppose that* w_n *converges weakly to* w. *Then* $w \in Au$.

Proof. By Lemma 1, we have that $\lim_{n\to\infty}(w_n, u_n-u)=0$. Then for each $v \in F$,

$$\liminf_{n\to\infty}(w_n, u_n-v)=\liminf_{n\to\infty}(w_n, u-v)=\lim_{n\to\infty}(w_n, u-v)=(w, u-v), \tag{3.1}$$

while, for each $v \in F$, there exists $w_v \in Au$ such that

$$(w_v, u-v) \leqq \liminf_{n\to\infty}(w_n, u_n-v). \tag{3.2}$$

Combining (3.1) with (3.2), we obtain that for any $v \in F$, there exists $w_v \in Au$ such that

$$(w_v, u-v) \leqq (w, u-v). \tag{3.3}$$

Suppose that $w \notin Au$. Then since Au is a closed convex set of F, there exists $z \in F$ such that

$$\langle w, z\rangle < \inf\{\langle y, z\rangle : y \in Au\}.$$

Put $v=u-z$ in (3.3). Then we obtain

$$\langle w_v, z\rangle \leqq \langle w, z\rangle < \langle w_v, z\rangle.$$

This is a contradiction. Thus we obtain that $w \in Au$.

LEMMA 2′. *Let* F *and* A *be as in Lemma* 2. *Let* $\{u_n\}$ *and* $\{w_n\}$ *be sequences such that* u_n *converges weakly to* u, $w_n \in Au_n$ *for each* $n \geqq 1$ *and* $\limsup_{n\to\infty}(w_n, u_n-u) \leqq 0$. *Suppose that* $\{w_n\}$ *is bounded in* F. *Then* $\bigcap_k \overline{\mathrm{co}}\{w_n : n \geqq k\} \subset Au$.

Proof. Let W be the set of all weak subsequential limits of $\{w_n\}$. Then by Lemma 2, $W \subset Au$. Since Au is closed and convex, $\overline{\mathrm{co}}\, W \subset Au$. Then since $\overline{\mathrm{co}}\, W = \bigcap_k \overline{\mathrm{co}}\{w_n : n \geqq k\}$ (cf. Bruck [4, Lemma 1.2]), the consequence follows.

LEMMA 3. *Let F and A be as in Lemma 2. Let L be a linear continuous and injective operator from F into itself and f be an element of F. Then the mapping L^*A_fL is a pseudo-monotone operator on F. Moreover if A maps bounded sets of F into bounded sets, then L^*A_fL is finitely continuous.*

Proof. Let $\{u_n\}$ and $\{w_n\}$ be sequences such that u_n converges weakly to u, $w_n \in L^*A_fLu_n$ for each $n \geqq 1$, and

$$\limsup_{n\to\infty}(w_n, u_n - u) \leqq 0.$$

Let $\{v_n\}$ be a sequence such that $v_n \in A(Lu_n + f)$ and $w_n = L^*v_n$, for each $n \geqq 1$. Then we have that for each $n \geqq 1$,

$$\liminf_{n\to\infty}(w_n, u_n - u) = \liminf_{n\to\infty}(v_n, Lu_n + f - (Lu + f)).$$

Then since A is pseudo-monotone and $Lu_n + f$ converges weakly to $Lu + f$, for given $z \in F$, there exists $v \in A(Lu + f)$ such that

$$\begin{aligned}(L^*v, u - z) &= (v, Lu + f - (Lz + f)) \\ &\leqq \liminf_{n\to\infty}(v_n, Lu_n + f - (Lz + f)) \\ &= \liminf_{n\to\infty}(v_n, Lu_n - Lz) \\ &= \liminf_{n\to\infty}(L^*v_n, u_n - z).\end{aligned}$$

Then since $L^*v \in L^*A(Lu + f)$, L^*A_fL is pseudo-monotone.

Suppose that A maps bounded sets of F into bounded sets. Then to see that L^*A_fL is finitely continuous, it is sufficient to show that L^*A_fL is demiclosed, i.e., the graph of L^*A_fL is strongly-weakly closed in $F \times F$, because $(L^*A_fL)G$ is relatively weakly compact for each compact subset G of F (cf. Browder [1, Proposition 2.6] or [2]). Let $\{u_n\}$ and $\{w_n\}$ be sequences in F such that $\{u_n\}$ is contained in the closed convex hull of a finite subset G of F, u_n converges strongly to u, $w_n \in L^*A_fLu_n$ and w_n converges weakly to w. Then since L^*A_fL is pseudo-monotone on F and $\limsup_{n\to\infty}\langle w_n, u_n - u\rangle \leqq 0$, we obtain, by Lemma 2, that $w \in L^*A_fLu$. This completes the proof.

LEMMA 4. *Let $A : L^2(0, T; H) \to L^2(0, T; H)$ be a multivalued operator with $D(A) = L^2(0, T; H)$. Let $a(t) \in L^1(0, T)$ and $f \in L^2(0, T; H)$. Then the equation*

$$u + L_aAu \ni f \tag{3.4}$$

has a solution u on $[0, T]$ if and only if the following equation has a solution v on $[0, T]$;

$$v + A(L_av + f) \ni 0. \tag{3.5}$$

Proof. Let u be a solution of (3.4). Put $z = u - f$. Then (3.4) is equal to

$$z + L_aA(z + f) \ni 0. \tag{3.6}$$

Then there exists $v \in L_a^{-1}z$ which satisfies (3.5). It is also easy to see the "if" part.

Proof of Theorem 1. For simplicity we write L instead of L_a. Let $f\in L^2(0,T;\ H)$. For each $n\geqq 1$, we set $L_n=(L+1/n)I$. Then since L is positive,

$$\langle L_n u,u\rangle=\langle L_n^* u,u\rangle\geqq\frac{1}{n}\|u\|^2 \quad \text{for } u\in L^2(0,T;\ H). \tag{3.7}$$

We first claim that for each $n\geqq 1$,

$$v+A(L_n v+f)\ni 0 \tag{3.8}$$

has a solution $v\in L^2(0,T;\ H)$. Since L_n^* is bijective by (3.7), $v\in L^2(0,T;\ H)$ is a solution of (3.8) if and only if v is a solution of

$$L_n^* v+L_n^* A(L_n v+f)\ni 0. \tag{3.9}$$

Then we show that (3.9) has a solution. For each $n\geqq 1$, $L_n^* A_f L_n$ is finitely continuous and pseudo-monotone by Lemma 3. Then $(1/n)I+L_n^* A_f L_n$ is also finitely continuous and pseudo-monotone for each $n\geqq 1$. From (2.4)′ we have that for each $u\in L^2(0,T;\ H)$ and $w\in A(L_n u+f)$,

$$\begin{aligned}\left\langle \frac{1}{n}u+L_n^* w,u\right\rangle&=\frac{1}{n}\|u\|^2+\langle w,L_n u\rangle\\ &=\frac{1}{n}\|u\|^2+\langle w,L_n u+f\rangle-\langle w,f\rangle \qquad (3.10)\\ &\geqq\frac{1}{n}\|u\|^2+\langle w,L_n u+f\rangle-\|f\|\left(c_1'+c_2'\left(\|a\|_{L^1(0,T)}\|u\|+\frac{1}{n}\|u\|+\|f\|\right)\right).\end{aligned}$$

Then by (2.5)′, (3.10) implies that for each $n\geqq 1$,

$$\lim_{\substack{z\in L_n^* A_f L_n u\\ \|u\|\to\infty}}\left\langle \frac{1}{n}u+z,u\right\rangle/\|u\|=\infty,$$

i.e., $(1/n)I+L_n^* A\mathscr{L}_n$ is coercive. By (2.4)′, $(1/n)I+L_n^* A_f L_n$ is bounded for each $n\geqq 1$. On the other hand, (3.7) implies that $L_n^*-(1/n)I$ is positive for each $n\geqq 1$. Then by [3, Thm. 7.8], we obtain that the sum of $(L_n^*-(1/n)I)$ and $((1/n)I+L_n^* A_f L_n)$ is surjective. Therefore (3.9) has a solution. Let $\{u_n\}\subset L^2(0,T;\ H)$ be a sequence such that $u_n+A(L_n u_n+f)\ni 0$, for each $n\geqq 1$. Let $\{w_n\}$ be a sequence such that $w_n\in A(L_n u_n+f)$ and $u_n+w_n=0$, for each $n\geqq 1$. Then by (2.4)′, we have that

$$\|u_n\|\leqq c_1'+c_2'\|L_n u_n+f\|\leqq c_1'+c_2'\left(\|a\|_{L^1(0,T)}\|u_n\|+\frac{1}{n}\|u_n\|+\|f\|\right). \tag{3.11}$$

Since $c_2'\|a\|_{L^1(0,T)}<1$, (3.11) implies that there exists $M>0$ such that

$$\|u_n\|\leqq M \quad \text{for all } n\geqq 1.$$

Then we can assume without any loss of generality that $u_n\rightharpoonup u$ and $w_n\rightharpoonup w$. Since $u_n+w_n=0$, we have that

$$\langle u_n,L_n u_n-Lu\rangle+\langle w_n,L_n u_n-Lu\rangle=0. \tag{3.12}$$

Since $\lim_{n\to\infty} L_n u = Lu$ and weak-$\lim_{n\to\infty} L_n u_n = $ weak-$\lim_{n\to\infty} Lu_n + (1/n)u_n = Lu$, it follows that

$$\begin{aligned}(3.13)\quad &\liminf_{n\to\infty} \langle u_n, L_n u_n - Lu\rangle \\ &= \liminf_{n\to\infty} \langle u_n - u, L_n u_n - L_n u\rangle + \lim_{n\to\infty} \langle u, L_n u_n - L_n u\rangle + \lim_{n\to\infty} \langle u_n, L_n u - Lu\rangle \\ &\geqq 0.\end{aligned}$$

Then combining (3.12) with (3.13), we have that

$$\limsup_{n\to\infty} \langle w_n, L_n u_n - Lu\rangle \leqq 0.$$

Then by Lemma 2, we obtain that $w \in A(Lu+f)$, and this implies $u + A(Lu+f) \ni 0$. Then by Lemma 4, (2.6) has a solution on $[0,T]$.

Proof of Corollary 1. We first note that from the condition (2.4), we have that for each $t \in [0,T]$,

$$\|y\|_{L^2(0,t;H)} \leqq \sqrt{T}\, c_1 + c_2 \|x\|_{L^2(0,t;H)},$$

for all $x \in L^2(0,t;\ H)$ and $y \in \tilde{A}x$. This implies that for each $t \in [0,T]$, (2.4)′ holds with $c_1' = \sqrt{T}\, c_1$ and $c_2' = c_2$. Then we choose $T' \in [0,T]$ such that $c_2 \|a\|_{L^1(0,T')} < 1$. Then by Theorem 1, there exists a solution u of (1.1) on $[0,T']$, i.e., there exists $w : [0,T'] \to H$ such that $w(t) \in Au(t)$ a.e. on $[0,T']$ and

$$u(t) + \int_0^t a(t-s) w(s)\, ds = f(t), \qquad 0 \leqq t \leqq T'.$$

By using the continuation argument employed in the proof of [1, Thm. 1], we show that $u(t)$ can be continued on $[0,T]$. Consider the equation

$$(3.14)\quad v(t) + \int_0^t a(t-s) Av(s)\, ds \ni f(t+T) - \int_0^T a(T+t-s) w(s)\, ds, \qquad 0 \leqq t \leqq T'',$$

where $T'' = \min(T', T-T')$. Let $f_1(t)$ denote the right side of (3.14). Then again by Theorem 1, there exists a solution v of (3.14). We define a function $\tilde{u} : [0, T'+T''] \ H$ by

$$\tilde{u}(t) = \begin{cases} u(t) & \text{on } [0,T'], \\ v(t-T') & \text{on } (T', T'+T'']. \end{cases}$$

Then $\tilde{u}$ is a solution of (1.1) on $[0, T+T'']$. By repeating this continuation, we obtain a solution u of (1.1) on $[0,T]$.

To prove Theorem 2, we need some lemmas and propositions. The continuation argument employed in Corollary 1 is available under the assumption of Theorem 2. Therefore in the following propositions, we assume that $c_2 \|a\|_{L^1(0,T)} < 1$ for $a(t) \in L^1(0,T)$ and A satisfying (2.4).

PROPOSITION 1. *Let $A : H \to H$ be an operator satisfying* (2.4) *and* (2.5). *Let $a(t) \in L^2(0,T)$ satisfy* (2.2) *and* (2.3). *Suppose that*

$$(3.15)\quad \begin{aligned}&\text{for each } f \in L^2(0,T;\ H),\ L_a^* \tilde{A}_f L_a \text{ is finitely continuous} \\ &\text{and pseudo-monotone on } L^2(0,T;\ H).\end{aligned}$$

Then for each $f \in L^2(0,T;\ H)$, (1.1) *has a solution on* $[0,T]$.

Proof. Let $f\in L^2(0,T;\ H)$. As in the proof of Theorem 1, we write L instead of L_a. From (2.4), (2.5) and (3.15), it follows that $((1/n)I+L^*\tilde{A}_fL)$ is finitely continuous, pseudo-monotone, coercive and bounded. Then since L^* is positive, by [3, Thm. 7.8], we obtain that for each $n\geqq 1$, there exists a solution u_n of the equation

$$L^*u_n+\frac{1}{n}u_n+L^*\tilde{A}_fLu_n\ni 0. \tag{3.16}$$

Since L^* is injective by (2.3), (3.16) implies

$$u_n+\frac{1}{n}L^{*-1}u_n+\tilde{A}_fLu_n\ni 0. \tag{3.17}$$

Multiplying (3.17) by u_n gives

$$\begin{aligned}\|u_n\|^2 &\leqq \|u_n\|\left(\sqrt{T}\,c_1+c_2\left(\|Lu_n+f\|\right)\right)\\ &\leqq \|u_n\|\left(\sqrt{T}\,c_1+c_2\|a\|_{L^1(0,T)}\|u_n\|+c_2\|f\|\right).\end{aligned} \tag{3.18}$$

Then it follows from (3.18) that there exists $M>0$ such that

$$\|u_n\|\leqq M \quad \text{for all } n\geqq 1. \tag{3.19}$$

Let $\{w_n\}$ be a sequence such that $w_n\in L^*\tilde{A}_fLu_n$ and $L^*u_n+(1/n)u_n+w_n=0$. Then by (3.19) and (2.4), we may assume that $u_n\to u$ and $w_n\to w$ in $L^2(0,T;\ H)$. Then since $(1/n)u_n\to 0$ and

$$\begin{aligned}\liminf_{n\to\infty}\left\langle L^*u_n,u_n-u\right\rangle &= \liminf_{n\to\infty}\left\langle L^*u_n-L^*u,u_n-u\right\rangle+\lim_{n\to\infty}\left\langle L^*u,u_n-u\right\rangle\\ &\geqq 0,\end{aligned}$$

multiplying (3.16) by u_n-u gives

$$\limsup_{n\to\infty}\left\langle w_n,u_n-u\right\rangle\leqq 0.$$

Then by Lemma 2, $w\in L^*\tilde{A}_fLu$ and therefore we obtain that $L^*u+L^*\tilde{A}(Lu+f)\ni 0$. Since L^* is injective, we have that $u+\tilde{A}(Lu+f)\ni 0$. Then by Lemma 4, (1.1) has a solution on $[0,T]$.

Theorem 2 follows from Proposition 1 and the following proposition.

PROPOSITION 2. *Let $A:H\to H$ be a finitely continuous pseudo-monotone operator on H. Let $a(t)\in L^2(0,T)$ satisfy* (2.2) *and* (2.3). *Then for each $f\in L^2(0,T;\ H)$, the operator $L^*\tilde{A}_fL$ is finitely continuous and pseudo-monotone on $L^2(0,T;\ H)$.*

For the sake of simplicity of the proof, we prove the case $f=0$ in Proposition 2. To prove Proposition 2, we need some lemmas. In the followings, we suppose that A and $a(t)$ satisfy the assumption of Proposition 2. We write L instead of L_a.

LEMMA 5. *Let $\{u_n\}$, $\{w_n\}\subset L^2(0,T;\ H)$ be sequences such that $u_n\to u$ in $L^2(0,T;\ H)$, $w_n\in\tilde{A}Lu_n$ for each $n\geqq 1$ and $\limsup_{n\to\infty}\langle w_n,Lu_n-Lu\rangle\leqq 0$. Then $\lim_{n\to\infty}\langle w_n,Lu_n-Lu\rangle=0$.*

Proof. We first observe that for each $t\in[0,T]$, $(Lu_n)(t)$ converges weakly to $(Lu)(t)$ in H. For each $z\in H$, $a(t)z\in L^2(0,T;\ H)$. Then since $u_n\to u$ in $L^2(0,T;\ H)$, we obtain that for each $t\in[0,T]$ and $z\in H$,

$$\begin{aligned}\lim_{n\to\infty}(z,(Lu_n)(t)-(Lu)(t))&=\lim_{n\to\infty}\left(z,\int_0^t a(t-s)(u_n(s)-u(s))\,ds\right)\\&=\lim_{n\to\infty}\int_0^t(a(t-s)z,\ u_n(s)-u(s))\,ds\\&=0.\end{aligned}$$

From the above observation and Lemma 1, we obtain that $\liminf_{n\to\infty}(w_n(t),(Lu_n)(t)-(Lu)(t))\geqq 0$ a.e. on $[0,T]$. We also have that there exists $M>0$ such that $|(Lu_n)(t)|\leqq M$ for all $n\geqq 1$ and $t\in[0,T]$. Then by the condition (2.4), we obtain that there exists $K>0$ such that $|(w_n(t),(Lu_n)(t)-(Lu)(t))|\leqq K$ for all $n\geqq 1$ and $t\in[0,T]$. The by Fatou's lemma, it follows that

$$\begin{aligned}0&\leqq\int_0^T\liminf_{n\to\infty}(w_n(t),(Lu_n)(t)-(Lu)(t))\,dt\\&\leqq\liminf_{n\to\infty}\langle w_n,Lu_n-Lu\rangle\\&\leqq\limsup_{n\to\infty}\langle w_n,Lu_n-Lu\rangle\\&\leqq 0,\end{aligned}$$

and this completes the proof.

LEMMA 6. *Let* $\{u_n\}$, $\{w_n\}\subset L^2(0,T;\ H)$ *be sequences such that* $u_n\to u$ *in* $L^2(0,T;\ H)$, $w_n\in\tilde{A}Lu_n$ *for all* $n\geqq 1$ *and* $\limsup_{n\to\infty}\langle w_n,Lu_n-Lu\rangle\leqq 0$. *Then for given* $\varepsilon>0$, *there exists* $n_0\geqq 1$ *such that* (*for all* $n\geqq n_0$),

$$(w_n(t),(Lu_n)(t)-Lu(t))>-\varepsilon,$$

for all $t\in[0,T]$.

Proof. Suppose that there exist $\varepsilon>0$, a sequence $\{t_i\}$ and a subsequence $\{u_{n_i}\}$ of $\{u_n\}$ such that

$$\lim_{i\to\infty}(w_{n_i}(t_i),(Lu_{n_i})(t_i)-(Lu)(t_i))\leqq-\varepsilon. \tag{3.20}$$

We may suppose without any loss of generality that $t_i\to t_0$. Then since $a(t)\in L^2(0,T)$ and $u_n\to u$ in $L^2(0,T;\ H)$, it is easy to see that $(Lu_{n_i})(t_i)$ converges weakly to $(Lu)(t_0)$ in H and $(Lu)(t_i)$ converges strongly to $(Lu)(t_0)$ in H. Then (3.20) implies

$$\lim_{i\to\infty}(w_{n_i}(t_i),(Lu_{n_i})(t_i)-(Lu)(t_0))\leqq-\varepsilon.$$

On the other hand, by Lemma 1, we have that

$$\liminf_{i\to\infty}(w_{n_i}(t_i),(Lu_{n_i})(t_i)-(Lu)(t_0))=0.$$

This is a contradiction.

LEMMA 7. *Let* $\{u_n\}$, $\{w_n\}$ *be sequences such that* $u_n\to u$ *in* $L^2(0,T;\ H)$, $w_n\in\tilde{A}Lu_n$ *for each* $n\geqq 1$ *and* $\limsup_{n\to\infty}\langle w_n,Lu_n-Lu\rangle\leqq 0$. *Then there exists a subsequence* $\{u_{n_i}\}$ *of* $\{u_n\}$ *such that*

$$\limsup_{i\to\infty}(w_{n_i}(t),(Lu_{n_i})(t)-(Lu)(t))\leqq 0\quad a.e.\ on\ [0,T]. \tag{3.21}$$

Proof. From Fatou's lemma, we have that

$$0 = \lim_{n\to\infty} \langle w_n, Lu_n - Lu\rangle \tag{3.22}$$
$$\leqq \int_0^T \limsup_{n\to\infty} (w_n(t), (Lu_n)(t) - (Lu)(t))\, dt$$
$$\leqq \lim_{k\to\infty} \int_0^T \sup_{n\geqq k} (w_n(t), (Lu_n)(t) - (Lu)(t))\, dt.$$

By Lemma 1, we have that

$$\liminf_{n\to\infty} (w_n(t), (Lu_n)(t) - (Lu)(t)) \geqq 0 \quad \text{a.e. on } [0,T]. \tag{3.23}$$

Then it is sufficient to prove that there exists a subsequence $\{u_{n_i}\}$ of $\{u_n\}$ such that

$$\lim_{j\to\infty} \int_0^T \sup_{i\geqq j} \left(w_{n_i}(t), (Lu_{n_i})(t) - (Lu)(t)\right) dt = 0. \tag{3.24}$$

In fact, if $\{u_{n_i}\}$ satisfies (3.24), we have, by (3.22), that

$$\int_0^T \limsup_{i\to\infty} \left(w_{n_i}(t), (Lu_{n_i})(t) - (Lu)(t)\right) dt = 0. \tag{3.25}$$

Then from (3.23) and (3.25), (3.21) follows. For each $n \geqq 1$ and $i \geqq 1$, we put

$$P_n(i) = \{t \in [0,T] : (w_n(t), (Lu_n)(t) - (Lu)(t)) \geqq 1/i\}.$$

We claim that there exists a subsequence $\{u_{n_i}\}$ of $\{u_n\}$ such that $m(P_{n_i}(i)) \leqq 1/2^i$ for all $i \geqq 1$, where m is the usual measure on $[0,T]$. Fix $i \geqq 1$ and choose $\varepsilon > 0$ such that $\varepsilon T + \varepsilon < (i2^i)^{-1}$. From Lemmas 5 and 6, we can choose n_i such that

$$\langle w_{n_i}, Lu_{n_i} - Lu\rangle < \varepsilon$$

and

$$\left(w_{n_i}(t), (Lu_{n_i})(t) - (Lu)(t)\right) \geqq -\varepsilon \quad \text{a.e. on } [0,T].$$

Then since

$$\varepsilon > \langle w_{n_i}, Lu_{n_i} - Lu\rangle$$
$$= \int_{P_{n_i}(i)} \left(w_{n_i}(t), (Lu_{n_i})(t) - (Lu)(t)\right) dt$$
$$+ \int_{[0,T]\setminus P_{n_i}(i)} \left(w_{n_i}(t), (Lu_{n_i})(t) - (Lu)(t)\right) dt$$
$$\geqq i^{-1} m\left(P_{n_i}(i)\right) - \varepsilon\left(T - m\left(P_{n_i}(i)\right)\right).$$

Then we obtain that $m(P_{n_i}(i)) < 2^{-i}$. Now we show that

$$\lim_{j\to\infty} \int_0^T \sup_{i\geqq j} \left(w_{n_i}(t), (Lu_{n_i})(t) - (Lu)(t)\right) dt = 0. \tag{3.26}$$

Since $\{u_n\}$ is bounded in $L^2(0,T;\ H)$, $a(t)\in L^2(0,T)$ and A satisfies (2.4), there exists $M>0$ such that $|(Lu_n)(t)|\leqq M$ and $|w_n(t)|\leqq M$ for all $n\geqq 1$ and $t\in[0,T]$. Then for given $\varepsilon>0$, we choose $i_0\geqq 1$ so large that $Ti_0^{-1}+2M^2(\Sigma_{i\geqq i_0}2^{-i})<\varepsilon$. Then we have

$$\begin{aligned}
&\int_0^T \sup_{i\geqq i_0}\left(w_{n_i}(t),\left(Lu_{n_i}\right)(t)-(Lu)(t)\right)dt\\
&\quad=\int_{\bigcup_{i\geqq i_0}P_{n_i}(i)} \sup_{i\geqq i_0}\left(w_{n_i}(t),\left(Lu_{n_i}\right)(t)-(Lu)(t)\right)dt\\
&\qquad+\int_{[0,T]\setminus\bigcup_{i\geqq i_0}P_{n_i}(i)} \sup_{i\geqq i_0}\left(w_{n_i}(t),\left(Lu_{n_i}\right)(t)-(Lu)(t)\right)dt\\
&\quad\leqq 2M^2\left(\sum_{i\geqq i_0}2^{-i}\right)+Ti_0^{-1}<\varepsilon.
\end{aligned}$$

Therefore we obtain (3.26). This completes the proof.

LEMMA 8. *Let* $\{u_n\}$, $\{w_n\}\subset L^2(0,T;\ H)$ *be sequences such that* $u_n\to u$ *in* $L^2(0,T;\ H)$, $w_n\in\tilde{A}Lu_n$ *for each* $n\geqq 1$ *and*

$$\limsup_{n\to\infty}\left(w_n(t),\left(Lu_n\right)(t)-(Lu)(t)\right)\leqq 0\quad a.e.\ on\ [0,T].$$

Suppose that $w_n\to w$ *in* $L^2(0,T;\ H)$. *Then* $w\in\tilde{A}Lu$.

Proof. It is easy to see that for each $t\in[0,T]$, $\{w_n(t)\}$ is bounded in H, because $(Lu_n)(t)$ is bounded in H and A satisfies (2.4), while, from the assumption, it is obvious that $w\in\overline{\mathrm{co}}\{w_n:n\geqq k\}$ for each $k\geqq 1$. Then we have that there exists a sequence $\{h_k\}$ such that $h_k\in\mathrm{co}\{w_n:n\geqq k\}$ and $h_k\to w$ in $L^2(0,T;\ H)$. Then we may assume without any loss of generality that $h_k(t)\to w(t)$ a.e. on $[0,T]$. Then since $h_k(t)\in\mathrm{co}\{w_n(t):n\geqq k\}$, we have that $w(t)\in\bigcap_k\mathrm{co}\{w_n(t):n\geqq k\}$ a.e. on $[0,T]$. Then by Lemma 2′, we obtain that $w(t)\in A(Lu)(t)$ a.e on $[0,T]$, i.e., $w\in\tilde{A}Lu$.

Remark. It is easy to verify that Lemmas 5, 6, 7, and 8 hold with $\tilde{A}$ replaced by $\tilde{A}_f$ ($f\in L^2(0,T;\ H)$.)

Proof of Proposition 2. For the sake of simplicity of the proof, we suppose that $f=0$. We prove that $L^*\tilde{A}L$ is pseudo-monotone on $L^2(0,T;\ H)$. Let $\{u_n\}$, $\{w_n\}\subset L^2(0,T;\ H)$ be sequences such that $u_n\to u$ in $L^2(0,T;\ H)$, $w_n\in\tilde{A}Lu_n$ for each $n\geqq 1$ and $\limsup_{n\to\infty}\langle w_n,Lu_n-Lu\rangle\leqq 0$. Let $v\in L^2(0,T;\ H)$. We show that there exists $w\in\tilde{A}Lu$ such that $\langle w,Lu-Lv\rangle\leqq\liminf_{n\to\infty}\langle w_n,Lu_n-Lv\rangle$. By Lemma 5, we have that $\lim_{n\to\infty}\langle w_n,Lu_n-Lu\rangle=0$. Hence we choose a subsequence $\{u_{n_i}\}$ of $\{u_n\}$ such that $w_{n_i}\to w$ in $L^2(0,T;\ H)$ and $\lim_{i\to\infty}\langle w_{n_i},Lu_{n_i}-Lv\rangle=\liminf_{n\to\infty}\langle w_n,Lu_n-Lv\rangle$. Then we have that

$$\begin{aligned}
\liminf_{n\to\infty}\langle w_n,Lu_n-Lv\rangle&=\lim_{i\to\infty}\langle w_{n_i},Lu_{n_i}-Lv\rangle\\
&=\lim_{i\to\infty}\langle w_{n_i},Lu_{n_i}-Lu\rangle+\lim_{i\to\infty}\langle w_{n_i},Lu-Lv\rangle\\
&=\lim_{i\to\infty}\langle w_{n_i},Lu-Lv\rangle\\
&=\langle w,Lu-Lv\rangle.
\end{aligned}$$

On the other hand, by Lemma 7, we may assume that

$$\limsup_{i\to\infty}\left\langle w_{n_i}(t),\,(Lu_{n_i})(t)-(Lu)(t)\right\rangle \leqq 0 \quad \text{a.e. on } [0,T].$$

Then by Lemma 8, it follows that $w\in\tilde{A}Lu$. Therefore $L^*\tilde{A}L$ is pseudo-monotone on $L^2(0,T;\ H)$. We can prove that $L^*\tilde{A}L$ is finitely continuous by the parallel argument as in the proof of Lemma 3. Then we omit the proof.

4. Examples.

Example 1. Let H be a real Hilbert space and let $A:H\to H$ be an operator defined by

$$Ax=\begin{cases}|\langle x_0,x\rangle|x/|x| & \text{if } x\neq 0,\\ \{x\in H:|x|\leqq 1.\} & \text{if } x=0,\end{cases}$$

where $x_0\in H$. Then it is easy to see that A is finitely continuous and pseudo-monotone on H. Let $a(t)$ be an element of $L^2(0,T)$ such that $a(0)>0$, a and a' are nonnegative convex on $[0,\infty)$. Then $a(t)$ satisfies (2.2) and (2.3) on each interval $[0,T](T>0)$. Moreover by Theorem 2, (1.1) has a solution on $[0,T]$ for every $f\in L^2(0,T;\ H)$.

Example 2. Let H be a real Hilbert space and $A:H\to H$ be a multivalued compact operator (i.e., A is upper semicontinuous and maps bounded sets to compact sets). Suppose that Au is convex for each $u\in H$. Then A is finitely continuous and pseudo-monotone on H. Suppose that A satisfies (2.4) and (2.5). Let $a(t)$ be as in example 1. Then for each $f\in L^2(0,T;\ H)$, (1.1) has a solution on $[0,T]$.

Acknowledgments. The author wishes to express his hearty thanks to Professor W. Takahashi, Professor J. A. Nohel and referees for many suggestions and advice regarding this paper.

REFERENCES

[1] V. BARBU, *Nonlinear Volterra equations in a Hilbert space*, this Journal, 6 (1975), pp. 728–741.
[2] F. E. BROWDER, *Fixed point theory of multivalued mappings in topological vector space*, Math. Ann., 177 (1968), pp. 283–301.
[3] ———, *Nonlinear operators and nonlinear equations of evolution in Banach spaces*, Proc. Symp. Pure Math., XVIII, 2, 1976.
[4] R. E. BRUCK, *On the almost-convergence of iterates of a nonexpansive mappings in Hilbert space and the structure of the weak w-limit set*, Israel J. Math., 29 (1978), pp. 1–16.
[5] M. G. CRANDALL AND J. A. NOHEL, *An abstract functional differential equation and a related nonlinear Volterra equation*, Israel J. Math., 29 (1978), pp. 313–328.
[6] G. GRIPENBERG, *An existence result for a nonlinear Volterra integral equation in a Hilbert space*, this Journal, 9 (1978), pp. 793–805.
[7] T. KIFFE AND M. STECHER, *L^2 solutions of Volterra integral equations*, this Journal, 10 (1979), pp. 274–280.
[8] ———, *Existence and uniqueness of solutions to abstract Volterra integral equations*, Proc. Amer. Math. Soc., 68 (1978), pp. 169–175.
[9] S. O. LONDON, *On an integral equation in a Hilbert space*, this Journal, 8 (1977), pp. 950–970.
[10] R. C. MACCAMY AND J. WONG, *Stability theorems for some functional equations*, Trans. Amer. Math. Soc., 164 (1972), pp. 1–37.
[11] J. A. NOHEL AND D. F. SHEA, *Frequency domain methods for Volterra equations*, Advances in Math., 3 (1976), pp. 278–304.
[12] O. STAFFANS, *Positive definite measures with applications a Volterra equation*, Trans. Amer. Math. Soc., 218 (1976), pp. 219–237.

SIAM J. MATH. ANAL.
Vol. 17, No. 2, March 1986

009

EIGENVALUES OF DIFFERENTIABLE POSITIVE DEFINITE KERNELS*

CHUNG-WEI HA†

Abstract. The main object of this note is to answer in the affirmative a conjecture in [4] that for an integral operator generated by a p times continuously differentiable positive definite kernel, the eigenvalues are $o(1/n^{p+1})$.

AMS(MOS) subject classifications. Primary 45B05, 45M05; secondary 47B10

Key words. L^2-kernel, singular value, eigenvalue, trace class operator

1. Introduction. A function $K(x,t)\in L^2[0,1]^2$ is known as an L^2-kernel. It defines a compact operator on $L^2[0,1]$ by

$$Kf(x)=\int_0^1 K(x,t)f(t)\,dt.$$

The adjoint K^* of K is generated by the kernel $\overline{K(t,x)}$ so that the operator K or its kernel $K(x,t)$ is Hermitian if $K(x,t)=\overline{K(t,x)}$ for almost all $0\leqq x,\ t\leqq 1$. Suppose that $K(x,t)$ is Hermitian and positive definite, that is,

$$\int_0^1\int_0^1 K(x,t)\overline{f(x)}f(t)\,dx\,dt\geqq 0$$

for all $f\in L^2[0,1]$. Then the spectrum of K consists of a sequence of positive eigenvalues $\{\lambda_n(K)\}$ which converges to 0. These eigenvalues are arranged in decreasing order and repeated according to their multiplicities. Assume also that $K(x,t)$ is continuous on $[0,1]^2$. If $\{\phi_n\}$ is the corresponding orthonormal sequence of eigenfunctions of K, then the well-known Mercer's theorem says that $K(x,t)$ has the eigenfunction expansion

$$K(x,t)=\sum_{n=1}^{\infty}\lambda_n(K)\phi_n(x)\overline{\phi_n(t)}, \tag{1}$$

where the series converges absolutely and uniformly on $[0,1]$. A kernel of this type generates an operator of trace class; the trace of K is given by

$$\sum_{n=1}^{\infty}\lambda_n(K)=\int_0^1 K(t,t)\,dt<\infty.$$

Consequently

$$\lambda_n(K)=o\left(\frac{1}{n}\right) \tag{2}$$

*Received by the editors January 17, 1984.
†Department of Mathematics, National Tsing Hua University, Hsin Chu, Taiwan, Republic of China.

as $n \to \infty$. Recently Reade [4] showed that if it is also assumed that $K(x,t) \in C^1[0,1]^2$, then (2) can be improved to $\lambda_n(K) = o(1/n^2)$ as $n \to \infty$. It is also suggested in [4] that if $p \geqq 1$ is an integer and if $K(x,t) \in C^p[0,1]^2$ is positive definite Hermitian, then

$$\lambda_n(K) = o\left(\frac{1}{n^{p+1}}\right)$$

as $n \to \infty$. The main object of this note is to prove this conjecture.

2. Singular values. Suppose that T is an integral operator not necessarily Hermitian. Then T^*T is positive definite Hermitian and so has a sequence of positive eigenvalues which converges to 0. The positive square roots of these eigenvalues, arranged in decreasing order and denoted by $\{s_n(T)\}$, are called singular values of T (see [3, Chap. 2]). It can be proved that T and T^* have the same singular values (see [3, p. 28]). If T is positive definite Hermitian, then the eigenvalues and the singular values of T coincide. The singular values of T have the characterization that

$$s_n(T) = \inf\{\|T|_H\|: H \text{ is a vector subspace, } \operatorname{codim} H \leqq n-1\},$$

where $T|_H$ denotes the restriction of T on H (see e.g. [1]). The infimum is attained when H is the vector subspace formed by $f \in L^2[0,1]$ which are orthogonal to the first $n-1$ eigenfunctions $\phi_1, \cdots, \phi_{n-1}$ of T^*T. Consequently, if $r \geqq 1$ is an integer and if H is a vector subspace of codimension $\leqq r$, then

$$s_{n+r}(T) \leqq s_n(T|_H). \tag{3}$$

Moreover, if K is positive definite Hermitian, then we can write $T^*KT = (K^{1/2}T)^*(K^{1/2}T)$, where $K^{1/2}$ denotes the positive square root of K, and so

$$\lambda_{m+n-1}(T^*KT) = s_{m+n-1}(T^*KT) \leqq s_m(K)s_n^2(T) = \lambda_m(K)\lambda_n(T^*T). \tag{4}$$

To prove the main results, we also need to compute the singular values of the operator

$$Jf(x) = \int_x^1 f(t)\,dt.$$

The numbers $s_n^2(J)$, which are eigenvalues of the operator J^*J, must satisfy

$$\phi_n(x) = \frac{1}{s_n^2(J)} \int_0^x \left\{ \int_t^1 \phi_n(s)\,ds \right\} dt,$$

that is,

$$\phi_n'' + \frac{1}{s_n^2(J)}\phi_n = 0, \qquad \phi_n(0) = \phi_n'(1) = 0.$$

A simple calculation shows (see [3, p. 120]) that

$$\lambda_n(J^*J) = s_n^2(J) = \frac{4}{\pi^2(2n-1)^2}. \tag{5}$$

3. The results. Suppose that the kernel $K(x,t)$ is positive definite Hermitian. If the symmetric derivative

$$K_{rr}(x,t) = \frac{\partial^{2r}}{\partial x^r \partial t^r} K(x,t) \tag{6}$$

exists and is continuous on $[0,1]^2$, then $K_{rr}(x,t)\in C[0,1]^2$ is also positive definite Hermitian, and so by Mercer's theorem, $K_{rr}(x,t)$ generates an integral operator of trace class. Indeed, it is shown in [2] that the eigenfunctions ϕ_n of K are in $C^r[0,1]$ and the eigenfunction expansion (1) can be differentiated term by term so that

$$K_{rr}(x,t)=\sum_{n=1}^{\infty}\lambda_n(K)\phi_n^{(r)}(x)\overline{\phi_n^{(r)}(t)}.$$

The series converges absolutely and uniformly on $[0,1]$.

THEOREM 1. *If $K(x,t)$ is positive definite Hermitian and if the symmetric derivative* (6) *exists and is continuous on* $[0,1]^2$, *then*

$$\sum_{n=1}^{\infty} n^{2r}\lambda_n(K)<\infty. \tag{7}$$

Consequently,

$$\lim_{n\to\infty} n^{2r+1}\lambda_n(K)=0. \tag{8}$$

Proof. The proof is suggested by a method of Krein (see [3, p. 121]). Let H_1 be the vector subspace formed by $f\in L^2[0,1]$ which are orthogonal to the constant function $e(t)\equiv 1$ and the function $K(t,0)$; then H_1 is of codimension $\leqq 2$. If $f\in H_1$, then

$$\int_0^1 f(t)\,dt=0 \quad\text{and}\quad \int_0^1 K(0,t)f(t)\,dt=0, \tag{9}$$

and so we have the expression

$$\begin{aligned} Kf(x)&=\int_0^1 K(x,t)f(t)\,dt \\ &=\int_0^x\left\{\int_0^1 K_{11}(y,s)\left(\int_s^1 f(t)\,dt\right)ds\right\}dy \\ &=J^*K_{11}Jf(x). \end{aligned} \tag{10}$$

We also have an expression symmetric to (10). Let G be the vector subspace formed by $g\in L^2[0,1]$ which are orthogonal to the functions $e(t)$ and $K(t,1)$. Then for $g\in G$

$$\int_0^1 g(t)\,dt=0 \quad\text{and}\quad \int_0^1 K(1,t)g(t)\,dt=0,$$

and so

$$\begin{aligned} Kg(x)&=\int_0^1 K(x,t)g(t)\,dt \\ &=\int_x^1\left\{\int_0^1 K_{11}(y,s)\left(\int_0^s g(t)\,dt\right)ds\right\}dy \\ &=JK_{11}J^*g(x). \end{aligned} \tag{11}$$

For $r=2$, let H_2 be the vector subspace formed by $f\in H_1$ which are orthogonal to the functions $J^*e(t)$ and $J^*K_{11}(t,1)$, then H_2 is codimension $\leqq 4$. If $f\in H_2$, then in addition to (9), f also satisfies

$$\int_0^1 Jf(t)\,dt=0 \quad \text{and} \quad \int_0^1 K_{11}(1,t)Jf(t)\,dt=0.$$

Now (9) holds, and by applying (11) to the function $Jf(t)$ with the kernel $K_{11}(x,t)$ in place of $K(x,t)$,

$$K_{11}Jf(x)=JK_{22}J^*Jf(x).$$

Substituting into (10), we have

$$Kf(x)=J^*JK_{22}J^*Jf(x).$$

If $r\geqq 3$, we can keep on applying (10) and (11) by turns. Let T_0 be the identity operator and for $1\leqq j\leqq r$, let

$$T_j=\underbrace{A\cdots J^*J}_{j \text{ factors}}$$

where $A=J$ if j is odd and $A=J^*$ if j is even; the operator T_j is the product of j factors which are either J or J^* appearing alternately. Let H_r be the vector subspace formed by $f\in L^2[0,1]$ which are orthogonal to the $2r$ functions

$$T_j^*e(t) \quad \text{and} \quad T_j^*K_{jj}(t,a)$$

for $j=0,1,\cdots,r-1$, where $a=0$ if j is even, $a=1$ if j is odd and $K_{00}(x,t)=K(x,t)$. Then H_r is of codimension $\leqq 2r$ and for $f\in H_r$,

$$Kf(x)=T_r^*K_{rr}T_rf(x).$$

Since $T_r^*T_r=(J^*J)^r$ and is positive definite Hermitian, $\lambda_n(T_r^*T_r)=\{\lambda_n(J^*J)\}^r$. By (3), (4) and (5) for $n\geqq 2r+1$

$$\begin{aligned}(12)\qquad \lambda_{2n}(K)&\leqq\lambda_{2n-1}(K)\leqq\lambda_{2n-2r-1}(T_r^*K_{rr}T_r)\\ &\leqq\lambda_{n-2r}(T_r^*T_r)\lambda_n(K_{rr})=\{\lambda_{n-2r}(J^*J)\}^r\lambda_n(K_{rr})\\ &\leqq\frac{4^r}{\pi^{2r}(2n-4r-1)^{2r}}\lambda_n(K_{rr}).\end{aligned}$$

As noted above, the operator K_{rr} is of trace class and so (7) is proved. (8) follows immediately from (7) (see [3, p. 122]).

THEOREM 2. *If $K(x,t)\in C^p[0,1]^2$ is positive definite Hermitian, then*

$$(13)\qquad \lambda_n(K)=o\left(\frac{1}{n^{p+1}}\right)$$

as $n\to\infty$.

Proof. If p is even, then (8) gives (13). If p is odd, then set $r=(p-1)/2$. Since $K_{rr}(x,t)\in C^1[0,1]^2$ is positive definite Hermitian, Reade's result [4] implies that

$$\lim_{n\to\infty} n^2\lambda_n(K_{rr})=0.$$

Relation (13) now follows from (12).

REFERENCES

[1] C.-W. HA, *Approximation numbers of linear operators and nuclear spaces*, J. Math. Anal. Appl., 46 (1974), pp. 292–311.

[2] T. T. KADOTA, *Term-by-term differentiability of Mercer's expansion*, Proc. Amer. Math. Soc., 18 (1967), pp. 69–72.

[3] I. C. GOHBERG AND M. G. KREIN, *Introduction to the Theory of Linear Nonselfadjoint Operators*, Transl. Math. Monographs, vol. 18, American Mathematical Society, Providence, RI, 1969.

[4] J. B. READE, *Eigenvalues of positive definite kernels*, this Journal, 14 (1983), pp. 152–157.

SIAM J. MATH. ANAL.
Vol. 17, No. 2, March 1986

010

POSITIVE DEFINITE C^p KERNELS*

J. B. READE†

Abstract. It was shown in [SIAM J. Math. Anal. 15 (1984), pp. 137–142] that the eigenvalues (λ_n) of any positive definite C^p kernel are $o(1/n^{p+1})$ as $n\to\infty$. More recently, Ha has proved [SIAM J. Math. Anal., 17 (1986), pp. 415–419] that, for p even, a stronger result holds, namely, that $\Sigma_1^\infty n^p\lambda_n<\infty$. It is the purpose of this note to show by counterexample that Ha's result is not true for p odd.

1. Introduction. Any p times continuously differentiable kernel $K(x,t)$ on $|x|\leqq 1$, $|t|\leqq 1$ gives rise to a compact operator

$$Tf(x)=\int_{-1}^{1} K(x,t)f(t)\,dt$$

on $L^2[-1,1]$. If $K(t,x)=\overline{K(x,t)}$, then T is also symmetric and so has an orthonormal sequence (ϕ_n) of eigenfunctions whose eigenvalues (λ_n) form a real sequence convergent to zero. Also $K(x,t)$ has the expansion

$$K(x,t)=\sum_1^\infty \lambda_n\phi_n(x)\overline{\phi_n(t)}$$

mean square convergent over $|x|\leqq 1$, $|t|\leqq 1$. If, further, $K(x,t)$ is positive definite, i.e.,

$$\int_{-1}^{1}\int_{-1}^{1} K(x,t)f(x)\overline{f(t)}dx\,dt\geqq 0$$

for all $f\in L^2[-1,1]$, then $\lambda_n\geqq 0$ for all n, and, if we assume

$$\lambda_1\geqq\lambda_2\geqq\cdots\geqq\lambda_n\geqq\cdots,$$

then we have $n^{p+1}\lambda_n\to 0$ as $n\to\infty$ if p odd, and $\Sigma_1^\infty n^p\lambda_n<\infty$ if p even. (See [1] (this issue, pp. 415–419), [2].)

We shall show both these results are best possible in the sense that, given any real decreasing sequence (λ_n) such that $n^{p+1}\lambda_n\to 0$ as $n\to\infty$ in case p odd, $\Sigma_1^\infty n^p\lambda_n<\infty$ in case p even, there exist positive definite C^p kernels whose eigenvalues are (λ_n). The divergence of the series $\Sigma_1^\infty 1/n\log n$ shows that any positive definite C^p kernel with eigenvalues $\lambda_n=1/n^{p+1}\log n$, where p is odd, cannot satisfy $\Sigma_1^\infty n^p\lambda_n<\infty$.

2. Trigonometric series.

LEMMA 1. *The trigonometric series $\Sigma_1^\infty n^{-1}\sin nt$ has uniformly bounded partial sums over all real t.*

Proof. The maximum value of $|\Sigma_1^N n^{-1}\sin nt|$ occurs at $t=\pi/N+1$ and is $\Sigma_1^N n^{-1}\sin n\pi/N+1$ which converges to $\int_0^\pi t^{-1}\sin t\,dt$ as $N\to\infty$. Q.E.D.

LEMMA 2. *If p is odd, and if (λ_n) is any decreasing real sequence such that $n^{p+1}\lambda_n\to 0$ as $n\to\infty$, then*

$$f(t)=\sum_1^\infty \lambda_n\cos nt$$

is C^p.

* Received by the editors April 3, 1984.
† Mathematics Department, Manchester University, Manchester, M13 9PL, England.

Proof. The series obtained by formally differentiating term by term p times is $\pm\sum_1^\infty n^p\lambda_n \sin nt$ which is uniformly convergent over all real t by Dirichlet's test (see [3, p. 50]) since $n^{p+1}\lambda_n$ decreases and converges to zero and $\sum_1^\infty n^{-1}\sin nt$ has uniformly bounded partial sums. Q.E.D.

LEMMA 3. *If p is even, and if (λ_n) is any real sequence such that $\sum_1^\infty n^p\lambda_n < \infty$, then*

$$f(t)=\sum_1^\infty \lambda_n \cos nt$$

is C^p.

Proof. p formal differentiations give $\pm\sum_1^\infty n^p\lambda_n \cos nt$ which is uniformly convergent by the Weierstrass M-test (see [3, p. 49]). Q.E.D.

3. The counterexamples. The sequence $(\cos \pi nt)_{n\geq 1}$ is orthonormal in $L^2[-1,1]$ and so, for any sequence (λ_n) such that $\sum_1^\infty \lambda_n^2 < \infty$, the kernel

$$K(x,t)=\sum_1^\infty \lambda_n \cos \pi nx \cos \pi nt = \frac{1}{2}\sum_1^\infty \lambda_n(\cos \pi n(x+t)+\cos \pi n(x-t))$$

is compact symmetric having eigenvalues (λ_n). $K(x,t)$ is C^p if, for p odd, λ_n decreases and $n^{p+1}\lambda_n \to 0$, or if, for p even, $\sum_1^\infty n^p\lambda_n < \infty$, by Lemmas 2 and 3.

REFERENCES

[1] C.-W. HA, *Eigenvalues of differentiable positive definite kernels*, this Journal, this issue, pp. 415–419.

[2] J. B. READE, *Eigenvalues of positive definite kernels* II, this Journal, 15 (1984), pp. 137–142.

[3] E. T. WHITTAKER AND G. N. WATSON, *A Course of Modern Analysis*, fourth edition, Cambridge University Press, Cambridge, 1927.

SIAM J. MATH. ANAL.
Vol. 17, No. 2, March 1986

011

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*

W. G. C. BOYD† AND T. M. DUNSTER†

Abstract. The asymptotic behaviour, as a parameter $u \to \infty$, of solutions of second-order linear differential equations with a turning point and a regular (double pole) singularity is considered. It is shown that the solutions can be approximated by expressions involving Bessel functions in a region which includes both the turning point and the singularity. Explicit error bounds for the difference between the approximations and the exact solutions are established. The theory is applied to find uniform asymptotic expansions for Legendre functions.

1. Introduction. In this paper we examine the asymptotic behaviour for large positive u of solutions of differential equations of the form

$$\frac{d^2w}{dx^2} = \{u^2 f(\alpha,x) + g(\alpha,x)\}w. \tag{1.1}$$

The equation is assumed to have a regular singularity at $x=0$ and, except when $\alpha=0$, a simple turning point at $x=x_t$. The real parameter α, which we define more precisely in a moment, is a measure of the severity of the singularity at $x=0$; we assume that it ranges in the interval $0 \leqq \alpha < \alpha_0$, where α_0 is a given constant which may be infinite. The position x_t of the turning point is assumed to be a continuous real function of α which tends to zero as $\alpha \to 0$. It is supposed that $g(\alpha,x)$ is small in absolute value compared with $u^2 f(\alpha,x)$ except near the turning point. We find asymptotic expansions and rigorous error bounds which are *uniformly* valid as α varies in an interval which includes $\alpha=0$. We consider separately the cases when the independent variable x takes real or complex values.

In the real-variable case, we shall suppose that the equation (1.1) is given in the x-interval (x_1,x_2) which includes 0, x_t; the endpoints x_1, x_2 may be infinite. Both $x^2 f(\alpha,x)$ and $x^2 g(\alpha,x)$ will be assumed to be infinitely differentiable functions of x and continuous functions of α and x simultaneously. (If one requires only asymptotic approximations to solutions of (1.1) instead of asymptotic expansions—i.e. requires only a finite number of terms—then the restriction to infinite differentiability can be relaxed.) The limits of $x^2 f(\alpha,x)$ and $x^2 g(\alpha,x)$ as $x \to 0$ will be assumed to be $\alpha^2/4$ and $-\frac{1}{4}$ respectively: if necessary, the parameter u is redefined to ensure that the condition on $g(\alpha,x)$ is met. We shall assume that for each nonzero value of α, $f(\alpha,x)$ has a simple zero at $x=x_t$ and elsewhere is nonzero. In the case $\alpha=0$, we shall assume that $f(0,x) \neq 0$ for $x \neq 0$ and that as $x \to 0$ the limit of $xf(0,x)$ is nonzero.

In the complex-variable case, we shall suppose that the equation (1.1) is given in a domain D of the complex z-plane which contains $z=0$ and $z=x_t$, and may be unbounded. Both $f(\alpha,z)$, $g(\alpha,z)$ will be assumed to be holomorphic functions of z and continuous functions of α and z simultaneously except when $z=0$. We shall assume that for each value of α the only singularities of $f(\alpha,z)$ and $g(\alpha,z)$ are at $z=0$: the

*Received by the editors May 24, 1983 and in final revised form August 9, 1984.
†School of Mathematics, University of Bristol, Bristol, England BS8 1TW.

Laurent series expansions of f and g about $z=0$ will be assumed to have leading terms $\alpha^2/4z^2$ and $-1/4z^2$ respectively. For each nonzero value of α, we shall assume that the function $f(\alpha,z)$ has a simple zero at the point on the real axis $z=x_t$ and elsewhere is nonzero. In the case $\alpha=0$, we shall assume that the function $f(0,z)$ is nonzero and has a simple pole at $z=0$.

In a review of unsolved problems in asymptotics, Olver (1975c, p. 117) suggests that uniform asymptotic expansions of solutions of (1.1) may be found in terms of Bessel functions of order $u\alpha$. We indeed find this to be the case. Expansions of this kind have been established previously by Olver (1958), (1974, Chap. 12) and Thorne (1957). Olver treats the case when α is small ($O(u^{-1})$). His 1974 results include explicit bounds for the error terms in the expansions. Thorne treats the case when α is not small. He does not obtain explicit error bounds. Our treatment differs in two major ways from those of Olver (1958) and Thorne (1957a). First, the expansions will be shown to be uniformly valid in an interval including $\alpha=0$ (so unifying the results of Olver and Thorne). Secondly, we obtain explicit error bounds, as Olver (1974) did for the case α small. The advantage of the first of these extensions is that the expansions can be used with confidence when α is either small or moderate (cf. Olver (1975a, pp. 138, 139)). The advantage of the second, other than the obvious computational consideration, is that the conditions under which the expansions may be regarded as asymptotic to exact solutions can readily be established a posteriori (cf. Olver (1975a, p. 139) and Olver (1980)).

We refer also to the work of Baldwin (1979). At first sight the problem he tackled may seem to be a special case of ours. This is not so: for example, in our notation the limit as $x\to 0$ of $xf(0,x)$ is zero for his problem. Nevertheless his work has some connections with ours.

We tackle the problem using the standard approach in problems of this sort: the comparison equation method. We transform the original equation into one of the form

$$\frac{d^2W}{d\zeta^2}=\left\{u^2\left(\frac{\alpha^2}{4\zeta^2}-\frac{1}{4\zeta}\right)+\frac{\psi(\alpha,\zeta)}{\zeta}-\frac{1}{4\zeta^2}\right\}W, \tag{1.2}$$

an equation which has the appropriate singular behaviour near $\zeta=0$ and a turning point at $\zeta=\alpha^2$. This choice of comparison equation is the same as that of Olver (1974, Chap. 12), but differs from that of Thorne (1957a). (Thorne's equation (2.4) may however be readily transformed to the above form.) With our choice of comparison equation, the transformation $z\to\zeta$ of the independent variable is analytic throughout D, whereas Thorne's is not at $z=0$.

The plan of the paper is as follows. In §2, we describe the formal series solutions for the real variable case and in §3 develop error bounds for them; §4 applies the results to the Legendre functions of real variable. In §5 we discuss the formal series and the corresponding error bounds for the complex variable case; §6 applies the result to the Legendre functions of complex argument.

2. Formal series solutions: the real-variable case. To transform the original equation (1.1) to the form (1.2) we define a new independent variable ζ and a new dependent variable W by

$$f(\alpha,x)\left(\frac{dx}{d\zeta}\right)^2=\frac{\alpha^2}{4\zeta^2}-\frac{1}{4\zeta}, \qquad \left(\frac{d\zeta}{dx}\right)^{1/2}w(x)=W(\zeta). \tag{2.1}$$

Without loss of generality we assume that $x_t > 0$ when $\alpha \neq 0$. We then specifically define ζ by

$$\text{(2.2a)} \qquad \int_{\alpha^2}^{\zeta} \frac{(\xi-\alpha^2)^{1/2}}{2\xi}\,d\xi = \int_{x_t}^{x} \{-f(\alpha,t)\}^{1/2}\,dt \qquad (x > x_t \text{ or } \zeta > \alpha^2),$$

$$\text{(2.2b)} \qquad \int_{\alpha^2}^{\zeta} \frac{(-\xi+\alpha^2)^{1/2}}{2\xi}\,d\xi = \int_{x_t}^{x} \frac{\{t^2 f(\alpha,t)\}^{1/2}}{t}\,dt \qquad (x < x_t \text{ or } \zeta < \alpha^2).$$

Each of the square roots in (2.2a, b) is taken to be positive or zero. When $\zeta < 0$ and $x < 0$, the integrals in (2.2b) are defined by their Cauchy principal values (for $\alpha \neq 0$). This latter convention introduces ambiguities into the transformation (2.2b) which we must resolve. Consider the left-hand side of (2.2b) when $\alpha \neq 0$: in any neighbourhood of $\zeta = 0$ there are pairs of values of ζ (one positive and the other negative) for which the integrals are equal; an analogous remark applies to the right-hand side of (2.2b). To remove the resulting ambiguities from the transformation $x \to \zeta$ we impose the restriction that the sign of ζ must be the same as that of the corresponding x.

We remark that the integrals on the left-hand sides (2.2a, b) are expressible in terms of elementary functions. They are respectively

$$\text{(2.3)} \qquad (\zeta-\alpha^2)^{1/2} - \alpha \tan^{-1}\frac{(\zeta-\alpha^2)^{1/2}}{\alpha}, \qquad \zeta > \alpha^2$$

and

$$(-\zeta+\alpha^2)^{1/2} - \frac{1}{2}\alpha \ln \frac{\alpha+(-\zeta+\alpha^2)^{1/2}}{|\alpha-(-\zeta+\alpha^2)^{1/2}|}, \qquad \zeta < \alpha^2.$$

The function $\zeta(x)$ defined by (2.2a, b) is monotonically increasing and infinitely differentiable in the interval $x_1 < x < x_2$. These results can be readily established by considering separately the intervals $(x_1, 0)$, $(0, x_t)$, (x_t, x_2) and neighbourhoods of the points $x = 0$, $x = x_t$. We denote the endpoints $\zeta(x_1)$ and $\zeta(x_2)$ by ζ_1, ζ_2 respectively.

The effect of the transformations (2.1) is to yield the new differential equation

$$\text{(2.4)} \qquad \frac{d^2W}{d\zeta^2} = \left\{u^2\left(\frac{\alpha^2}{4\zeta^2} - \frac{1}{4\zeta}\right) + \frac{\psi(\alpha,\zeta)}{\zeta} - \frac{1}{4\zeta^2}\right\} W,$$

where, with $'$ representing differentiation with respect to ζ,

$$\text{(2.5)} \qquad \frac{\psi(\alpha,\zeta)}{\zeta} = x'^{1/2}\frac{d^2}{d\zeta^2}(x'^{-1/2}) + g x'^2 + \frac{1}{4\zeta^2}.$$

A straightforward calculation yields the result

$$\text{(2.6)} \quad \psi(\alpha,\zeta) = \frac{1}{16}(\zeta-\alpha^2)^{-2}(\zeta+4\alpha^2) + \frac{1}{64}\zeta^{-1}(\zeta-\alpha^2)f^{-3}(5\dot{f}^2 - 4f\ddot{f} - 16gf^2),$$

where $\cdot$ represents differentiation with respect to x. Since $x(\zeta)$ and $x^2 g(x)$ are infinitely differentiable and $x'(\zeta) > 0$, it follows from (2.5) that $\psi(\alpha,\zeta)$ is an infinitely differentiable function of ζ for $\zeta_1 < \zeta < \zeta_2$, except possibly at $\zeta = 0$ where the third and fourth terms on the right-hand side of (2.5) have singularities of $O(\zeta^{-2})$. Consideration of the Maclaurin series of $x^2 g(\alpha, x)$ about $x = 0$ however, shows that as $\zeta \to 0$, the third and fourth terms together are $O(\zeta^{-1})$ and so the Maclaurin series of $\psi(\alpha,\zeta)$ about $\zeta = 0$ exists to all orders; in particular $\psi(\alpha,\zeta)$ is infinitely differentiable at $\zeta = 0$.

There is a further property of $\psi(\alpha,\zeta)$ that we shall require: continuity as a function of α, ζ simultaneously. This result is needed to ensure that the asymptotic expansions we find are *uniformly* valid near $\zeta=0$, $\alpha=0$. For convenience we introduce the notation

$$f(\alpha,x)=\frac{x_t(\alpha)-x}{4x^2}p(\alpha,x).$$

We may then, following Olver (1975a, p. 142), prove the following result.

LEMMA 1. *Assume that*

(i) $p(\alpha,x)$, $\partial p(\alpha,x)/\partial x$, $\partial^2 p(\alpha,x)/\partial x^2$ *and* $g(\alpha,x)$ *are continuous functions of* α *and* x *in the region* $0\leqq\alpha<\alpha_0$, $x_1<x<x_2$;

(ii) $p(\alpha,x)$ *is positive throughout the same region*;

(iii) $|\partial^3 p(\alpha,x)/\partial x^3|$ *is bounded in a neighbourhood of the point* $\alpha=x=0$;

(iv) $x_t(\alpha)$ *is a continuous function of* α *when* $0\leqq\alpha<\alpha_0$, *which takes positive values when* $\alpha>0$ *and tends to zero as* $\alpha\to 0$.

Then the function $\psi(\alpha,\zeta)$ *defined by* (2.5) *is continuous in the corresponding region of the* (α,ζ)*-plane.*

There is a corresponding result for complex x, which is fairly easy to prove. The real-variable proof is rather more difficult, but largely follows that given by Olver (1975a, p. 142). It is outlined in Appendix A.

If the term $\psi(\alpha,\zeta)/\zeta$ is neglected in (2.4) the resulting equation has solutions of the form $\zeta^{1/2}\mathscr{C}_{u\alpha}(u\zeta^{1/2})$ when $\zeta>0$, where $\mathscr{C}$ denotes the Bessel functions J, Y, or any combination of them; when $\zeta<0$, solutions are $|\zeta|^{1/2}\mathscr{L}_{u\alpha}(u|\zeta|^{1/2})$, where $\mathscr{L}$ denotes the modified Bessel functions I, K, or any combination of them. The solutions corresponding to J, Y and to I, K each constitute linearly independent pairs. For the purposes of our error analysis we shall need to select solutions which satisfy the more demanding restriction that they constitute numerically satisfactory pairs (Olver (1974, p. 154)). The solutions corresponding to J, Y and to I, K do each constitute numerically satisfactory pairs in $\zeta>0$ and $\zeta<0$ respectively. When ζ is complex, neither pair is satisfactory.

When account is taken of the term $\psi(\alpha,\zeta)/\zeta$ in (2.4), the functions we have just discussed are only the first terms in asymptotic expansions of solutions of (2.4): we seek expansions of the form

$$\zeta^{1/2}\mathscr{C}_{u\alpha}(u\zeta^{1/2})\sum_{s=0}^{\infty}\frac{A_s(\alpha,\zeta)}{u^{2s}}+\frac{\zeta}{u}\mathscr{C}'_{u\alpha}(u\zeta^{1/2})\sum_{s=0}^{\infty}\frac{B_s(\alpha,\zeta)}{u^{2s}}\qquad(\zeta>0),\tag{2.7}$$

where $\mathscr{C}$ denotes J, Y or any combination of these. Formal substitution of this series into (2.4) yields

$$\zeta^{1/2}\mathscr{C}_{u\alpha}(u\zeta^{1/2})\sum_{s=0}^{\infty}\frac{C_s}{u^{2s}}+u\zeta\mathscr{C}'_{u\alpha}(u\zeta^{1/2})\sum_{s=0}^{\infty}\frac{D_s}{u^{2s}}=0,$$

where

$$C_s=\zeta^{-1}A'_s+A''_s-\zeta^{-1}\psi A_s-\frac{1}{2}\zeta^{-1}B_s-B'_s-\alpha^2\zeta^{-1}B'_s,$$

$$D_s=\zeta^{-1}A'_s+\zeta^{-1}B'_{s-1}+B''_{s-1}-\zeta^{-1}\psi B_{s-1}.$$

In these and subsequent expressions, quantities with negative indices are to be understood to be identically zero. The differential equation can then be formally satisfied provided $C_s=0$, $D_s=0$ for $s=0,1,2,\cdots$. Thus we require

$$|\zeta-\alpha^2|^{1/2}\left(|\zeta-\alpha^2|^{1/2}B_s\right)'=\zeta A_s''+A_s'-\psi A_s, \tag{2.8a}$$

$$A_s'=-\left(\zeta B_{s-1}'\right)'+\psi B_{s-1}. \tag{2.8b}$$

These relations can be satisfied by

$$B_s(\alpha,\zeta)=|\zeta-\alpha^2|^{-1/2}\int_{\alpha^2}^{\zeta}|\xi-\alpha^2|^{-1/2}\big(\xi A_s''(\alpha,\xi)+A_s'(\alpha,\xi) -\psi(\alpha,\xi)A_s(\alpha,\xi)\big)\,d\xi, \tag{2.9a}$$

$$A_s(\alpha,\zeta)=-\zeta B_{s-1}'(\alpha,\zeta)+\int_{\alpha^2}^{\zeta}\psi(\alpha,\xi)B_{s-1}(\alpha,\xi)\,d\xi+\lambda_s. \tag{2.9b}$$

In (2.9b), λ_s is an arbitrary constant of integration; the corresponding constant in (2.9a) has been set equal to zero to ensure that $B_s(\zeta)$ is differentiable at $\zeta=\alpha^2$. Without loss of generality, we take $\lambda_0=1$, so that $A_0(\alpha,\zeta)\equiv 1$. Relations (2.9a, b) then successively determine $B_0,A_1,B_1,A_2,\cdots$. The remaining arbitrary constants $\lambda_1,\lambda_2,\cdots$ may be assigned values by reference to the specific properties of that solution whose asymptotic expansion is being found. It follows immediately from (2.9a, b) that $A_s(\alpha,\zeta)$, $B_s(\alpha,\zeta)$ are infinitely differentiable functions of ζ in the interval $0\leqq\zeta<\zeta_2$ for each value of α (cf. Olver (1974, p. 410)).

When $\zeta<0$, we seek formal series solutions of the form

$$|\zeta|^{1/2}\mathscr{L}_{u\alpha}\left(u|\zeta|^{1/2}\right)\sum_{s=0}^{\infty}\frac{A_s(\zeta)}{u^{2s}}+\frac{|\zeta|}{u}\mathscr{L}_{u\alpha}\left(u|\zeta|^{1/2}\right)\sum_{s=0}^{\infty}\frac{B_s(\zeta)}{u^{2s}} \tag{2.10}$$

where $\mathscr{L}$ denotes I, K or any combination of these. One finds that the coefficients $A_s(\zeta)$, $B_s(\zeta)$ in (2.10) satisfy the same differential equations (2.8a, b) as do the coefficients in the formal series solution (2.7): we define $A_s(\zeta)$, $B_s(\zeta)$ for $\zeta<0$ by analytic continuation from $\zeta<0$, so that (2.9a, b) are valid in $\zeta<0$ too. (Thus $A_s(\zeta)$, $B_s(\zeta)$ are infinitely differentiable in $\zeta_1<\zeta<\zeta_2$.)

3. Error bounds (real variables). To verify that the series (2.7), (2.10) are uniformly valid asymptotic expansions of an exact solution $W(u,\alpha,\zeta)$ of (2.4) we terminate the series after a finite number of terms and then find an upper bound for the error. Thus when $\zeta>0$, we write

$$W(u,\alpha,\zeta)=\zeta^{1/2}\mathscr{C}_{u\alpha}\left(u\zeta^{1/2}\right)\sum_{s=0}^{n}\frac{A_s(\alpha,\zeta)}{u^{2s}}+\frac{\zeta}{u}\mathscr{C}'_{u\alpha}\left(u\zeta^{1/2}\right)\sum_{s=0}^{n-1}\frac{B_s(\alpha,\zeta)}{u^{2s}}+\varepsilon_{2n+1}(u,\alpha,\zeta), \tag{3.1}$$

where $\mathscr{C}$ denotes J, Y or a combination of them. The standard method of obtaining a bound for ε_{2n+1} is first to find a differential equation for ε_{2n+1}; then this differential equation is re-expressed as a Volterra integral equation, a bound for the solution of which may be found by the method of successive approximations (Olver (1974, p. 141)).

The differential equation for ε_{2n+1} is readily found to be

$$(3.2)\quad \varepsilon''_{2n+1}+\left\{-u^2\left(\frac{\alpha^2}{4\zeta^2}-\frac{1}{4\zeta}\right)+\frac{1}{4\zeta^2}\right\}\varepsilon_{2n+1}$$
$$=-\frac{1}{u^{2n}}\left|\zeta-\alpha^2\right|^{1/2}\left\{\left|\zeta-\alpha^2\right|^{1/2}B_n(\alpha,\zeta)\right\}'\zeta^{-1/2}\mathscr{C}_{u\alpha}(u\zeta^{1/2})+\frac{\psi}{\zeta}\varepsilon_{2n+1},$$

where the primes denote ζ-derivatives. Given some arbitrary value $\tilde{\zeta}$ $(0\leqq\tilde{\zeta}\leqq\zeta_2)$, consider that solution of (3.2) for which $\varepsilon_{2n+1}(u,\alpha,\tilde{\zeta})=\varepsilon'_{2n+1}(u,\alpha,\tilde{\zeta})=0$: except possibly when $\tilde{\zeta}=0$ or ∞, it satisfies the integral equation

$$(3.3)\quad \varepsilon_{2n+1}(u,\alpha,\zeta)=\int_{\tilde{\zeta}}^{\zeta}K(\zeta,\xi)\left[-\frac{1}{u^{2n}}\left\{\left|\xi-\alpha^2\right|^{1/2}B_n(\alpha,\xi)\right\}'\xi^{1/2}\mathscr{C}_{u\alpha}(u\xi^{1/2})\right.$$
$$\left.+\left|\xi-\alpha^2\right|^{-1/2}\psi(\alpha,\xi)\varepsilon_{2n+1}(u,\alpha,\xi)\right]d\xi$$

where

$$K(\zeta,\xi)=\pi\left|\xi-\alpha^2\right|^{1/2}\left\{\zeta^{1/2}Y_{u\alpha}(u\zeta^{1/2})\xi^{-1/2}J_{u\alpha}(u\xi^{1/2})\right.$$
$$\left.-\zeta^{1/2}J_{u\alpha}(u\zeta^{1/2})\xi^{-1/2}Y_{u\alpha}(u\xi^{1/2})\right\}.$$

Bounds for ε_{2n+1} and ε'_{2n+1} can now be found by using a standard theorem given by Olver (1974, p. 219).

To use this theorem, we introduce a modulus function $M_\nu(x)$, a phase function $\theta_\nu(x)$, and a weight function $E_\nu(x)$ which are related by

$$E_\nu(x)J_\nu(x)=M_\nu(x)\cos\theta_\nu(x),\qquad E_\nu^{-1}(x)Y_\nu(x)=M_\nu(x)\sin\theta_\nu(x).$$

Following Olver (1974, p. 437) we define $E_\nu(x)$ as follows. Let $x=X_\nu$ be the smallest positive root of the equation

$$(3.4)\qquad J_\nu(x)=-Y_\nu(x).$$

Then define

$$E_\nu(x)=\begin{cases}\left(-Y_\nu(x)/J_\nu(x)\right)^{1/2}, & 0<x\leqq X_\nu,\\ 1, & x\geqq X_\nu.\end{cases}$$

As Olver shows, $E_\nu(x)$ is a monotonically nonincreasing function of x. One finds

$$M_\nu(x)=\begin{cases}\left(-2J_\nu(x)Y_\nu(x)\right)^{1/2}, & 0<x\leqq X_\nu,\\ \left(J_\nu^2(x)+Y_\nu^2(x)\right)^{1/2}, & x\geqq X_\nu,\end{cases}$$

$$\theta_\nu(x)=\begin{cases}-\pi/4, & 0<x\leqq X_\nu,\\ \tan^{-1}\left(Y_\nu(x)/J_\nu(x)\right), & x\geqq X_\nu.\end{cases}$$

The branch of the inverse tangent is chosen so that $\theta_\nu(x)$ is continuous. We also introduce modulus and phase functions for $J_\nu'(x)$, $Y_\nu'(x)$ as follows. Define

$$E_\nu(x)J_\nu'(x)=N_\nu(x)\cos\omega_\nu(x), \qquad E_\nu^{-1}(x)Y_\nu'(x)=N_\nu(x)\sin\omega_\nu(x),$$

where $E_\nu(x)$ is unchanged. Thus

$$N_\nu(x)=\begin{cases}\left(\dfrac{Y_\nu^2(x)J_\nu'^2(x)+J_\nu^2(x)Y_\nu'^2(x)}{-J_\nu(x)Y_\nu(x)}\right)^{1/2}, & 0<x\leqq X_\nu,\\ \left(J_\nu'^2(x)+Y_\nu'^2(x)\right)^{1/2}, & x\geqq X_\nu,\end{cases}$$

$$\omega_\nu(x)=\begin{cases}\tan^{-1}\left(-\dfrac{J_\nu(x)Y_\nu'(x)}{Y_\nu(x)J_\nu'(x)}\right), & 0<x\leqq X_\nu,\\ \tan^{-1}\left(\dfrac{Y_\nu'(x)}{J_\nu'(x)}\right), & x\geqq X_\nu.\end{cases}$$

The branches of the inverse tangents are chosen so that $\omega_\nu(x)$ is continuous, and $\omega_\nu(x)\to-\pi/4$ as $x\to0$ ($\nu>0$) or $\omega_\nu(x)\to-\pi/2$ as $x\to0$ ($\nu=0$).

Before stating the theorem on error bounds, we introduce the following constants:

$$\begin{aligned}\kappa^+&=\sup\left\{\pi|x^2-\nu^2|^{1/2}M_\nu^2(x)\right\},\\ \mu_1^+&=\sup\left\{\pi|x^2-\nu^2|^{1/2}E_\nu(x)M_\nu(x)|J_\nu(x)|\right\},\\ \mu_2^+&=\sup\left\{\pi|x^2-\nu^2|^{1/2}E_\nu^{-1}(x)M_\nu(x)|Y_\nu(x)|\right\},\end{aligned}\tag{3.5}$$

each supremum being evaluated over $x>0$ and $\nu\geq0$. In Appendix B we show that κ^+ exists and is finite. The proofs for μ_1^+ and μ_2^+ are similar. Numerical calculations indicate that $\kappa^+=2.08\cdots$ with the supremum being achieved as $x,\nu\to\infty$ such that $x\sim\nu-(1.33\cdots)^{1/3}\nu^{1/3}$ (see part (d) of Lemma 2, Appendix B), and that $\mu_1^+=\mu_2^+=2$ with the supremum being achieved as $x\to\infty$ for fixed ν.

THEOREM 1. *With the conditions described in* §§1 *and* 2, *equation* (2.4) *has, for each pair of values of* u *and* α *and each nonnegative integer* n, *solutions* $W_{2n+1,1}(u,\alpha,\zeta)$, $W_{2n+1,2}(u,\alpha,\zeta)$ *which are infinitely differentiable in* $0<\zeta<\zeta_2$ *and satisfy*

$$W_{2n+1,1}(u,\alpha,\zeta)=\zeta^{1/2}J_{u\alpha}(u\zeta^{1/2})\sum_{s=0}^{n}\frac{A_s(\alpha,\zeta)}{u^{2s}}+\frac{\zeta}{u}J_{u\alpha}'(u\zeta^{1/2})\sum_{s=0}^{n-1}\frac{B_s(\alpha,\zeta)}{u^{2s}}+\varepsilon_{2n+1,1}(u,\alpha,\zeta),\tag{3.6}$$

$$W_{2n+1,2}(u,\alpha,\zeta)=\zeta^{1/2}Y_{u\alpha}(u\zeta^{1/2})\sum_{s=0}^{n}\frac{A_s(\alpha,\zeta)}{u^{2s}}+\frac{\zeta}{u}Y_{u\alpha}'(u\zeta^{1/2})\sum_{s=0}^{n-1}\frac{B_s(\alpha,\zeta)}{u^{2s}}+\varepsilon_{2n+1,2}(u,\alpha,\zeta),\tag{3.7}$$

where

(3.8)

$$\frac{|\varepsilon_{2n+1,1}(u,\alpha,\zeta)|}{\zeta^{1/2}M_{u\alpha}(u\zeta^{1/2})}, \qquad \frac{|\partial\varepsilon_{2n+1,1}(u,\alpha,\zeta)/\partial\zeta|}{(1/2)uN_{u\alpha}(u\zeta^{1/2})+(1/2)\zeta^{-1/2}M_{u\alpha}(u\zeta^{1/2})}$$

$$\leqq \frac{1}{u^{2n+1}}\mu_1^+ E_{u\alpha}^{-1}(u\zeta^{1/2})\mathscr{V}_{0,\zeta}\left\{|\xi-\alpha^2|^{1/2}B_n(\xi)\right\}\exp\left(\frac{\kappa^+}{u}\mathscr{V}_{0,\zeta}\left\{|\xi-\alpha^2|^{1/2}B_0(\xi)\right\}\right),$$

(3.9)

$$\frac{|\varepsilon_{2n+1,2}(u,\alpha,\zeta)|}{\zeta^{1/2}M_{u\alpha}(u\zeta^{1/2})}, \qquad \frac{|\partial\varepsilon_{2n+1,2}(u,\alpha,\zeta)/\partial\zeta|}{(1/2)uN_{u\alpha}(u\zeta^{1/2})+(1/2)\zeta^{-1/2}M_{u\alpha}(u\zeta^{1/2})}$$

$$\leqq \frac{1}{u^{2n+1}}\mu_2^+ E_{u\alpha}(u\zeta^{1/2})\mathscr{V}_{\zeta_2,\zeta}\left\{|\xi-\alpha^2|^{1/2}B_n(\xi)\right\}\exp\left(\frac{\kappa^+}{u}\mathscr{V}_{\zeta_2,\zeta}\left\{|\xi-\alpha^2|^{1/2}B_0(\xi)\right\}\right).$$

Here and subsequently, we use the symbol $\mathscr{V}$ to denote the variational operator, a formal definition of which may be found in Olver (1974, p. 27).

These results are proved following the standard method referenced below equation (3.3). An essential feature of the proofs is the observation that

$$E_{u\alpha}(u\zeta^{1/2}) \leqq E_{u\alpha}(u\xi^{1/2}) \quad \text{for } 0<\xi<\zeta. \tag{3.10}$$

In the notation of (3.3), choose $\tilde{\zeta}=0$, $\mathscr{C}\equiv J$ to arrive at (3.6), (3.8) and $\tilde{\zeta}=\zeta_2$, $\mathscr{C}\equiv Y$ to arrive at (3.7), (3.9). When $\zeta_2=\infty$, (3.9) is meaningful provided the variations of $(\zeta-\alpha^2)^{1/2}B_s(\zeta)$ converge at infinity. A sufficient condition for this is that as $\zeta\to\infty$, the ζ-derivatives $\psi^{(s)}(\alpha,\zeta)$ are $O(\zeta^{-1/2-s-\sigma})$ where σ is some positive constant (cf. Olver (1974, p. 445, exercise 4.2)). If the variations fail to converge, $\tilde{\zeta}$ has to be chosen to be finite, $\tilde{\zeta}_0$ say, and the results of the theorem then apply to the interval $0<\zeta<\tilde{\zeta}_0$.

The bounds (3.8), (3.9) can be used to deduce the asymptotic nature of the expansions (2.7). Consider (3.8). In the first place, we see at once that

$$\varepsilon_{2n+1,1}(u,\alpha,\zeta)=\zeta^{1/2}J_{u\alpha}(u\zeta^{1/2})O(\zeta) \quad \text{as } \zeta\to 0 \tag{3.11}$$

from which we deduce that the exact solution $W_{2n+1,1}(u,\alpha,\zeta)$ is recessive as $\zeta\to 0$. To within a multiplicative constant, therefore, $W_{2n+1,1}(u,\alpha,\zeta)$ is uniquely defined. Now consider the asymptotic nature of the approximation as $u\to\infty$. Since the variations, $\mathscr{V}_{0,\zeta}$, that occur in (3.8) are bounded for each α, we deduce

$$\varepsilon_{2n+1,1}(u,\alpha,\zeta)=\zeta^{1/2}E_{u\alpha}^{-1}(u\zeta^{1/2})M_{u\alpha}(u\zeta^{1/2})O(u^{-2n-1}) \quad \text{as } u\to\infty$$

uniformly in the ζ-interval $(0,\zeta_2)$, provided $\zeta_2<\infty$, and the α-interval $[0,\alpha_0']$, where α_0' is any constant satisfying $0<\alpha_0'<\alpha_0$. (The uniformity is a consequence of Lemma 1.) This result can be used to show that (2.7), with $\mathscr{C}=J$, is a uniformly valid compound asymptotic expansion of $W_{2n+1,1}(u,\alpha,\zeta)$. To do so, express

$$E_\nu^{-1}(x)M_\nu(x)=a_\nu(x)J_\nu(x)+b_\nu(x)xJ_\nu'(x),$$

where

$$a_\nu(x)=\begin{cases}2^{1/2}, & 0<x\leqq X_\nu,\\ -\frac{1}{2}\pi x Y_\nu'(x)M_\nu(x), & x>X_\nu,\end{cases}$$

and

$$b_\nu(x)=\begin{cases}0, & 0<x\leqq X_\nu,\\ \frac{1}{2}\pi Y_\nu(x)M_\nu(x), & x>X_\nu.\end{cases}$$

It then suffices to observe that $a_\nu(x)$ and $b_\nu(x)$ are uniformly bounded for $\nu\geqq 0$, $x>0$. (This can be demonstrated from considerations analogous to those discussed in the proof of Lemma 2 in Appendix B.)

The other solution, $W_{2n+1,2}(u,\alpha,\zeta)$, is dominant as $\zeta\to 0$: it must be identified by reference to its properties elsewhere (e.g. as $\zeta\to\zeta_2$). Proceeding as above, one can show that (2.7), with $\mathscr{C}=Y$, is a compound asymptotic expansion of $W_{2n+1,2}(u,\alpha,\zeta)$, uniformly valid as $u\to\infty$ in the ζ-interval $(0,\zeta_2)$ and the α-interval $[0,\alpha_0']$.

There are two further remarks to be made about (3.6), (3.7). The first is that there are solutions $W_1(u,\alpha,\zeta)$ and $W_2(u,\alpha,\zeta)$ which are independent of n and have the infinite series (2.7), with $\mathscr{C}=J$, Y respectively, as their compound asymptotic expansions. (This can be shown by the method of Olver (1974, Chap. 10, §6).) The second is that when $\zeta_2=\infty$, the expansions will be *uniformly* valid in the ζ-interval $(0,\infty)$ provided the variations of $(\zeta-\alpha^2)^{1/2}B_s(\zeta)$ converge at infinity; sufficient conditions for this to be true have been already noted.

When $\zeta<0$ the observation that $I_{u\alpha}(u|\zeta|^{1/2})$ and $K_{u\alpha}(u|\zeta|^{1/2})$ are respectively monotonically increasing and decreasing functions of $|\zeta|$ for fixed u, α leads to the following result.

THEOREM 2. *With the conditions described in* §§1 *and* 2, *equation* (2.4) *has, for each pair of values of* u *and* α *and each nonnegative integer* n, *solutions* $W_{2n+1,3}(u,\alpha,\zeta)$, $W_{2n+1,4}(u,\alpha,\zeta)$ *which are infinitely differentiable in* $\zeta_1<\zeta<0$ *and satisfy*

$$W_{2n+1,3}(u,\alpha,\zeta)=|\zeta|^{1/2}I_{u\alpha}\big(u|\zeta|^{1/2}\big)\sum_{s=0}^{n}\frac{A_s(\alpha,\zeta)}{u^{2s}}+\frac{|\zeta|}{u}I'_{u\alpha}\big(u|\zeta|^{1/2}\big)\sum_{s=0}^{n-1}\frac{B_s(\alpha,\zeta)}{u^{2s}}+\varepsilon_{2n+1,3}(u,\alpha,\zeta), \tag{3.12}$$

$$W_{2n+1,4}(u,\alpha,\zeta)=|\zeta|^{1/2}K_{u\alpha}\big(u|\zeta|^{1/2}\big)\sum_{s=0}^{n}\frac{A_s(\alpha,\zeta)}{u^{2s}}+\frac{|\zeta|}{u}K'_{u\alpha}\big(u|\zeta|^{1/2}\big)\sum_{s=0}^{n-1}\frac{B_s(\alpha,\zeta)}{u^{2s}}+\varepsilon_{2n+1,4}(u,\alpha,\zeta), \tag{3.13}$$

where

$$\frac{|\varepsilon_{2n+1,3}(u,\alpha,\zeta)|}{|\zeta|^{1/2}I_{u\alpha}\big(u|\zeta|^{1/2}\big)},\quad \frac{|\partial\varepsilon_{2n+1,3}(u,\alpha,\zeta)/\partial\zeta|}{(1/2)(1+u\alpha)|\zeta|^{-1/2}I_{u\alpha}\big(u|\zeta|^{1/2}\big)+uK^{-1}_{u\alpha}\big(u|\zeta|^{1/2}\big)K_{u\alpha+1}\big(u|\zeta|^{1/2}\big)I_{u\alpha}\big(u|\zeta|^{1/2}\big)} \tag{3.14}$$

$$\leqq\frac{1}{u^{2n+1}}\kappa^{-}\mathscr{V}_{0,\zeta}\big\{|\xi-\alpha^2|^{1/2}B_n(\xi)\big\}\exp\Big(\frac{\kappa^{-}}{u}\mathscr{V}_{0,\zeta}\big\{|\xi-\alpha^2|^{1/2}B_0(\xi)\big\}\Big),$$

$$\frac{|\varepsilon_{2n+1,4}(u,\alpha,\zeta)|}{|\zeta|^{1/2}K_{u\alpha}\big(u|\zeta|^{1/2}\big)},\quad \frac{|\partial\varepsilon_{2n+1,4}(u,\alpha,\zeta)/\partial\zeta|}{(1/2)(1+u\alpha)|\zeta|^{-1/2}K_{u\alpha}\big(u|\zeta|^{1/2}\big)+uK_{u\alpha+1}\big(u|\zeta|^{1/2}\big)} \tag{3.15}$$

$$\leqq\frac{1}{u^{2n+1}}\kappa^{-}\mathscr{V}_{\zeta_1,\zeta}\big\{|\xi-\alpha^2|^{1/2}B_n(\xi)\big\}\exp\Big(\frac{\kappa^{-}}{u}\mathscr{V}_{\zeta_1,\zeta}\big\{|\xi-\alpha^2|^{1/2}B_0(\xi)\big\}\Big).$$

Here,

$$\kappa^- = \sup\left\{2(x^2+\nu^2)^{1/2} I_\nu(x) K_\nu(x)\right\}, \tag{3.16}$$

the supremum being evaluated over $x>0$ and $\nu \geqq 0$. The existence of the supremum is discussed in Appendix B. Numerical calculations indicate $\kappa^- = 1.0667\cdots$, the supremum being achieved when $x = 1.07501\cdots$, $\nu = 0$.

The bounds (3.14), (3.15) can be used to deduce the asymptotic nature of the expansions (2.10); the discussion is similar to that which follows Theorem 1. Thus the solutions $W_{2n+1,3}(u,\alpha,\zeta)$ and $W_{2n+1,4}(u,\alpha,\zeta)$ are respectively recessive and dominant at $\zeta = 0$. With $\mathscr{L} = I$, K, (2.10) is a uniformly valid compound asymptotic expansion as $u \to \infty$ of $W_{2n+1,3}(u,\alpha,\zeta)$ and $W_{2n+1,4}(u,\alpha,\zeta)$ respectively. There are solutions $W_3(u,\alpha,\zeta)$, $W_4(u,\alpha,\zeta)$ independent of n for which (2.10), with $\mathscr{L} = I$, K, respectively, is a compound asymptotic expansion. When $\zeta_1 = -\infty$, the theorem will be meaningful and the expansions will be uniformly valid in the ζ-interval $(-\infty, 0)$ if as $\zeta \to -\infty$, the ζ-derivatives $\psi^{(s)}(\alpha,\zeta)$ are $O(|\zeta|^{-1/2-s-\sigma})$ for some positive constant σ.

4. Legendre functions of large order and degree: real variables. We shall illustrate the results of the previous two sections by constructing asymptotic expansions for $\nu \to \infty$ of solutions for the associated Legendre equation which are uniformly valid for $x \geqq 0$, $0 \leqq \mu/(\nu+\frac{1}{2}) \leqq 1-\delta$ for arbitrarily small $\delta > 0$. (See Olver 1975c, p. 125).) Corresponding results for negative values of ν, μ, or x may be obtained by using the appropriate connection formulae. Our choice of solutions is $P_\nu^{-\mu}(x)$, $Q_\nu^\mu(x)$ in $x>1$ and the Ferrers functions $\mathrm{P}_\nu^{-\mu}(x)$, $\mathrm{Q}_\nu^\mu(x)$ in $x<1$; our notation is that of Olver (1974, Chap. 5, definitions (12.04), (13.14), (15.01), (15.02)) respectively.

The functions $(x^2-1)^{1/2}P_\nu^{-\mu}(x)$, $(x^2-1)^{1/2}Q_\nu^\mu(x)$, $(1-x^2)^{1/2}\mathrm{P}_\nu^{-\mu}(x)$, $(1-x^2)^{1/2}\mathrm{Q}_\nu^\mu(x)$ each satisfy

$$\frac{d^2w}{dx^2} = \left\{\frac{\mu^2-1}{(1-x^2)^2} - \frac{\nu(\nu+1)}{1-x^2}\right\} w. \tag{4.1}$$

To within multiplicative constants, these solutions may be identified by their properties that $\mathrm{P}_\nu^{-\mu}(x)$, $P_\nu^{-\mu}(x)$ are recessive at $x=1$ for $\mu \geqq 0$ and $Q_\nu^\mu(x)$, $\mathrm{Q}_\nu^\mu(x)$ are recessive at $x=\infty$ for $\nu \geqq -\frac{1}{2}$. In terms of our previous notation, we define

$$u = \nu + \frac{1}{2}, \quad \alpha = \frac{\mu}{(\nu+1/2)}, \quad f(x) = \frac{\alpha^2}{(1-x^2)^2} - \frac{1}{1-x^2}, \quad g(x) = \frac{-1}{(1-x^2)^2} + \frac{1}{4(1-x^2)}.$$

The singular point is at $x=1$ and the turning point at $x_t = (1-\alpha^2)^{1/2}$. In order that the turning point be real, it is necessary for us to demand $\alpha < 1$: this is the reason for our requiring $\mu < \nu + \frac{1}{2}$.

The x-ζ transformations (2.2a, b) are, for this problem,

$$\int_{\alpha^2}^{\zeta} \frac{(\xi-\alpha^2)^{1/2}}{2\xi}\,d\xi = -\int_{(1-\alpha^2)^{1/2}}^{x} \frac{(1-\alpha^2-t^2)^{1/2}}{1-t^2}\,dt \tag{4.2a}$$

$$\left(\zeta > \alpha^2 \text{ or } x < (1-\alpha^2)^{1/2}\right),$$

$$\int_{\alpha^2}^{\zeta} \frac{(\alpha^2-\xi)^{1/2}}{2\xi}\,d\xi = -\int_{(1-\alpha^2)^{1/2}}^{x} \frac{(t^2-1+\alpha^2)^{1/2}}{1-t^2}\,dt \tag{4.2b}$$

$$\left(\zeta < \alpha^2 \text{ or } x > (1-\alpha^2)^{1/2}\right),$$

the latter integrals being understood to be the Cauchy principal values in $\zeta<0$ or $x>1$ (for $\alpha\neq 0$). The integrals on the right-hand side of (4.2a, b) may be evaluated. For example (4.2b) is

$$\ln\frac{x+(x^2-1+\alpha^2)^{1/2}}{(1-\alpha^2)^{1/2}}+\frac{\alpha}{2}\ln\frac{|1-x|}{1+x}\,\frac{x+1-\alpha^2-\alpha(x^2-1+\alpha^2)^{1/2}}{x-1+\alpha^2+\alpha(x^2-1+\alpha^2)^{1/2}} \tag{4.3}$$

(see, for example, Gradshteyn and Ryzhik (1980, p. 81)). With the new dependent variable W defined by

$$w=\left(\frac{\zeta-\alpha^2}{1-\alpha^2-x^2}\right)^{1/4}\left(\frac{1-x^2}{2\zeta}\right)^{1/2}W$$

one finds that the differential equation (4.1) transforms to

$$\frac{d^2W}{d\zeta^2}=\left\{u^2\left(\frac{\alpha^2}{4\zeta^2}-\frac{1}{4\zeta}\right)-\frac{1}{4\zeta^2}+\frac{\psi(\alpha,\zeta)}{\zeta}\right\}W, \tag{4.4}$$

where

(4.5)

$$\psi(\alpha,\zeta)=\frac{1}{16}(\zeta-\alpha^2)^{-2}\left[\zeta+4\alpha^2+\frac{1-x^2}{\zeta}\left(\frac{\zeta-\alpha^2}{1-\alpha^2-x^2}\right)^3\{(-4\alpha^2+1)x^2+(\alpha^4-1)\}\right].$$

Before proceeding to establish the asymptotic nature of the solutions of (4.4), it is convenient at this stage to describe the nature of the x-ζ transformation near $x=1$ and $x=\infty$. As $x\to 1$, one finds that $\zeta\to 0$ such that

$$1-x=\frac{e^2}{2(1+\alpha)^{1+1/\alpha}(1-\alpha)^{1-1/\alpha}}\zeta+O(\zeta^2). \tag{4.6}$$

We remark that this result is uniformly valid for all values of α (see proof of Lemma 1 in Appendix A); when $\alpha=0$ the limiting form of (4.6), $1-x\sim\frac{1}{2}\zeta$, applies. As $x\to\infty$, one finds $\zeta\to-\infty$ such that

$$x=\frac{1}{2}(1+\alpha)^{1/2+\alpha/2}(1-\alpha)^{1/2-\alpha/2}\exp|\zeta|^{1/2}\left\{1+O\left(|\zeta|^{-1/2}\right)\right\}. \tag{4.7}$$

These results follow readily from (2.3) and (4.3).

We shall now find uniform asymptotic expansions for $\mathrm{P}_\nu^{-\mu}(x)$ in $0\leqq x<1$ and for $P_\nu^{-\mu}(x)$ in $x>1$. Both solutions are recessive as $x\to 1$. In particular

$$\mathrm{P}_\nu^{-\mu}(x)=\frac{(1-x)^{\mu/2}}{2^{\mu/2}\Gamma(\mu+1)}(1+o(1))\quad\text{as } x\to 1 \tag{4.8}$$

(Olver (1974, p. 186)). It follows from (3.11) that for each nonnegative integer n,

$$\mathrm{P}_\nu^{-\mu}(x)=c_{2n+1,1}\left(\frac{\zeta-\alpha^2}{1-\alpha^2-x^2}\right)^{1/4}\zeta^{-1/2}W_{2n+1,1}(u,\alpha,\zeta). \tag{4.9}$$

for some constant $c_{2n+1,1}$. The coefficient $c_{2n+1,1}$ can be evaluated by comparing (4.8) with the behaviour of $W_{2n+1,1}(u,\alpha,\zeta)$ as $\zeta\to 0$. The behaviour of $J_{u\alpha}(u\zeta^{1/2})$ near $\zeta=0$ implies that

$$(4.10)\quad W_{2n+1,1}(u,\alpha,\zeta)=\frac{u^{u\alpha}}{2^{u\alpha}\Gamma(1+u\alpha)}\zeta^{1/2+u\alpha/2}\cdot\left(1+\alpha\sum_{s=0}^{n-1}\frac{B_s(\alpha,0)}{(\nu+1/2)^{2s+1}}+\sum_{s=1}^{n}\frac{A_s(\alpha,0)}{(\nu+1/2)^{2s}}\right)\quad\text{as }\zeta\to 0,$$

where $A_s(\alpha,\zeta)$, $B_s(\alpha,\zeta)$ are defined by (2.9a, b). We shall not attempt to define the arbitrary constants λ_s which appear in (2.9b) at this stage but, following Thorne (1957b), determine them recursively (see (6.8)). It follows from (4.6), (4.8)–(4.10) that

$$(4.11)\quad c_{2n+1,1}=e^{\mu}\left(\nu+\frac{1}{2}+\mu\right)^{-\nu/2-1/4-\mu/2}\left(\nu+\frac{1}{2}-\mu\right)^{\nu/2+1/4-\mu/2}\cdot\left(1+\alpha\sum_{s=0}^{n-1}\frac{B_s(\alpha,0)}{(\nu+1/2)^{2s+1}}+\sum_{s=1}^{n}\frac{A_s(\alpha,0)}{(\nu+1/2)^{2s}}\right)^{-1}.$$

We subsequently derive another expression for $c_{2n+1,1}$ (equation (6.10)).

A powerful check on the validity of (4.9) is provided by evaluating left- and right-hand sides at $x=0$. The left-hand side yields (Olver (1974, p. 187))

$$(4.12)\quad \mathrm{P}_{\nu}^{-\mu}(0)=2^{-\mu}\pi^{-1/2}\cos\left(\tfrac{1}{2}\pi\nu-\tfrac{1}{2}\pi\mu\right)\Gamma\left(\tfrac{1}{2}\nu+\tfrac{1}{2}-\tfrac{1}{2}\mu\right)/\Gamma\left(\tfrac{1}{2}\nu+1+\tfrac{1}{2}\mu\right).$$

By using the Liouville–Green expansion of $J_{\mu}((\nu+\frac{1}{2})\zeta^{1/2})$ for large argument and order, one can show that to leading order the right-hand side of (4.9) yields, for $n=0$,

$$(4.13)\quad c_{1,1}(1-\alpha^2)^{-1/4}\left(\frac{2}{\pi(\nu+1/2)}\right)^{1/2}\cos\left(\tfrac{1}{2}\pi\nu-\tfrac{1}{2}\pi\mu\right).$$

When Stirling's formula is used in (4.12), it is found that to leading order, (4.12) and (4.13) are indeed the same. This provides independent verification of the correctness of the asymptotic theory.

In the interval $x>1$, one can show in a similar manner that, at least pointwise,

$$(4.14)\quad P_{\nu}^{-\mu}(x)=c_{2n+1,3}\left(\frac{\zeta-\alpha^2}{1-\alpha^2-x^2}\right)^{1/4}|\zeta|^{-1/2}W_{2n+1,3}(u,\alpha,\zeta),$$

where $c_{2n+1,3}=c_{2n+1,1}$. Actually, it can be shown that (4.14) is *uniformly* valid in $x>1$: it follows from (4.5)–(4.7) that $\psi^{(s)}(\alpha,\zeta)=O(|\zeta|^{-s-1})$ as $\zeta\to-\infty$, and as we remarked in §3, this guarantees uniform validity.

Next, we find a uniform asymptotic expansion for the function $Q_{\nu}^{\mu}(x)$ in $x>1$. Its distinguishing property is that it is recessive as $x\to\infty$. In applying Theorem 2 to the function $Q_{\nu}^{\mu}(x)$ we first observe that since $\psi^{(s)}(\alpha,\zeta)=O(|\zeta|^{-s-1})$ as $\zeta\to-\infty$, the theorem is meaningful for $\zeta_1=-\infty$, and the expansions are uniformly valid in $\zeta<0$. Since $I_{u\alpha}(u|\zeta|^{1/2})$, $K_{u\alpha}(u|\zeta|^{1/2})$ are respectively dominant and recessive as $\zeta\to-\infty$, we deduce that

$$(4.15)\quad Q_{\nu}^{\mu}(x)=c_{2n+1,4}\left(\frac{\zeta-\alpha^2}{1-\alpha^2-x^2}\right)^{1/4}|\zeta|^{-1/2}W_{2n+1,4}(u,\alpha,\zeta),\qquad x>1$$

for some constant $c_{2n+1,4}$. We evaluate $c_{2n+1,4}$ by considering the behaviour of left- and right-hand sides of (4.15) as $x\to 1$. For the left-hand side we have (Olver (1974, Chap. 5, (12.21) and (13.14)))

$$Q_\nu^\mu(x)=e^{\mu\pi i}2^{\mu/2-1}\Gamma(\mu)(x-1)^{-\mu/2}(1+o(1)) \quad \text{as } x\to 1. \tag{4.16}$$

The behaviour of $K_{u\alpha}(u|\zeta|^{1/2})$ near $\zeta=0$ implies that on the right-hand side,

(4.17)

$$W_{2n+1,4}(u,\alpha,\zeta)=\frac{2^{u\alpha-1}\Gamma(u\alpha)}{u^{u\alpha}}|\zeta|^{1/2-u\alpha/2}$$
$$\cdot\left(1-\alpha\sum_{s=0}^{n-1}\frac{B_s(\alpha,0)}{(\nu+1/2)^{2s+1}}+\sum_{s=1}^{n}\frac{A_s(\alpha,0)}{(\nu+1/2)^{2s}}+\delta_{2n+1,4}+o(1)\right) \quad \text{as } \zeta\to 0,$$

for $\alpha>0$, where $\delta_{2n+1,4}$ is a constant bounded by

$$|\delta_{2n+1,4}|\leqq\frac{1}{u^{2n+1}}\kappa^-\mathscr{V}_{0,-\infty}\left\{|\xi-\alpha^2|^{1/2}B_n(\xi)\right\}\exp\left(\frac{\kappa^-}{u}\mathscr{V}_{0,-\infty}\left\{|\xi-\alpha^2|^{1/2}B_0(\xi)\right\}\right).$$

It follows from (4.6), (4.15)–(4.17) that

$$c_{2n+1,4}=e^{-\mu+\mu\pi i}\left(\nu+\frac{1}{2}+\mu\right)^{\nu/2+1/4+\mu/2}\left(\nu+\frac{1}{2}-\mu\right)^{-\nu/2-1/4+\mu/2} \tag{4.18}$$
$$\cdot\left(1-\alpha\sum_{s=0}^{n-1}\frac{B_s(\alpha,0)}{(\nu+1/2)^{2s+1}}+\sum_{s=1}^{n}\frac{A_s(\alpha,0)}{(\nu+1/2)^{2s}}+\delta_{2n+1,4}\right)^{-1}.$$

A continuity argument shows that (4.18) is appropriate for $\alpha=0$ also. We subsequently derive another expression for $c_{2n+1,4}$ (equation (6.11)).

It remains to find a uniform asymptotic expansion for $\mathrm{Q}_\nu^{-\mu}(x)$ in $0\leqq x<1$. (The result is simpler than that for $\mathrm{Q}_\nu^\mu(x)$.) One could arrive at a result following the method described by Olver (1974, Chap. 12) in dealing with the corresponding problem when μ is not large: effectively this means taking $\zeta_2=\zeta(0)$ in (3.9). As Olver remarks, this method has disadvantages, and we shall leave the derivation of the result to the complex variable theory of §6 (see (6.12)).

Our results for the Legendre functions, here and in §6, are more general than those given by Olver (1974, Chap. 12) or Thorne (1957b). They may be regarded as complementary to the results of Olver (1975b), which are uniformly valid in the interval $-1<x<1$.

5. Expansions for complex argument. The definition of the x-ζ transformation given in (2.2a, b) is no longer appropriate when x is supposed to be a complex variable. Now we replace x by z and define

$$\int_{\alpha^2}^{\zeta}\frac{(\xi-\alpha^2)^{1/2}}{2\xi}d\xi=\int_{x_t}^{z}\frac{(t^2f(t))^{1/2}}{t}dt \tag{5.1}$$

where account has to be taken of the branches resulting both from the square roots and the logarithmic singularities. It is easy to see that the branches can be chosen so that $\zeta(z)$ is analytic at both $z=0$ and $z=x_t$; subject to this requirement, the actual choice of branches is unimportant. The dependent variable is transformed as in the real case:

$$\left(\frac{d\zeta}{dz}\right)^{1/2}w(z)=W(\zeta). \tag{5.2}$$

With the transformations (5.1), (5.2) we arrive at (2.4) again, i.e.

$$\frac{d^2W}{d\zeta^2} = \left\{ u^2\left(\frac{\alpha^2}{4\zeta^2} - \frac{1}{4\zeta} \right) + \frac{\psi(\alpha,\zeta)}{\zeta} - \frac{1}{4\zeta^2} \right\} W. \tag{5.3}$$

We continue to assume the parameters satisfy $\alpha \geqq 0$, $u > 0$, but now suppose ζ is a complex variable and that (5.3) holds in some complex domain Δ which includes $\zeta = 0$ and $\zeta = \alpha^2$ and in which $\psi(\alpha,\zeta)$ is holomorphic.

If the term $\psi(\alpha,\zeta)/\zeta$ is neglected in (5.3) the resulting equation has solutions of the form $\zeta^{1/2}\mathscr{C}_{u\alpha}(u\zeta^{1/2})$ where $\mathscr{C}$ denotes one of the Bessel functions J, $H^{(1)}$, $H^{(2)}$, or any combination of them. We shall seek formal series solutions of (5.3) of the form

$$\zeta^{1/2}\mathscr{C}_{u\alpha}(u\zeta^{1/2}) \sum_{s=0}^{\infty} \frac{A_s(\alpha,\zeta)}{u^{2s}} + \frac{\zeta}{u}\mathscr{C}'_{u\alpha}(u\zeta^{1/2}) \sum_{s=0}^{\infty} \frac{B_s(\alpha,\zeta)}{u^{2s}}, \tag{5.4}$$

where the coefficients $A_s(\alpha,\zeta)$, $B_s(\alpha,\zeta)$ are the analytic continuations of those defined by (2.9a, b). The square root $\zeta^{1/2}$ which occurs in (5.4) will be defined by

$$\chi_0 < \arg\zeta^{1/2} < \chi_0 + \pi \tag{5.5}$$

for some χ_0 ($-\pi \leqq \chi_0 \leqq 0$). The choice of χ_0 is determined by the particular problem under consideration: a common choice is $\chi_0 = 0$. Let $\boldsymbol{\Delta}$ denote the intersection of Δ and the ζ-plane cut along $\arg\zeta = 2\chi_0$: we shall seek solutions of the form (5.4) in the domain $\boldsymbol{\Delta}$.

It is convenient for our purposes to introduce the notation $\mathscr{C}^{(j)}$, where $\mathscr{C}^{(j)}$ denotes $2J$, $H^{(1)}$, $H^{(2)}$ for $j = 0, 1, 2$ respectively. Subsequently, we shall suppose j is enumerated modulo 3. See also Baldwin (1979).

In order that we may construct auxiliary functions with the appropriate asymptotic properties we have to take account of the asymptotic behaviour of the functions $\mathscr{C}_\nu^{(j)}(z)$ in the complex z-plane. It might therefore seem natural to define weight functions $E_\nu^{(j)}(z)$ by

$$\left| \exp\left(-i\int_\nu^z \frac{(t^2-\nu^2)^{1/2}}{t}\, dt \right) \right|. \tag{5.6}$$

However this choice proves inappropriate when ν and z are both small and would not give us the *uniformly* valid result we seek. Instead we define weight functions which are close to (5.6) when ν or $|z|$ is large but otherwise differ significantly.

To this end, we first note that (Olver (1974, p. 438))

$$X_\nu = \nu - c\left(\frac{1}{2}\nu\right)^{1/3} + O(\nu^{-1/3}) \quad \text{as } \nu \to \infty,$$

where c denotes the negative root of the equation $\mathrm{Ai}(x) = \mathrm{Bi}(x)$ of smallest absolute value. Let us define $\hat{\nu}$ to be real (nonnegative) solution of

$$\nu = \hat{\nu} - c\left(\frac{1}{2}\hat{\nu}\right)^{1/3}.$$

Then $\hat{X}_\nu$, the smallest zero of $J_{\hat{\nu}}(x) + Y_{\hat{\nu}}(x)$ (cf. (3.4)), is such that

$$\hat{X}_\nu = \nu + O(\nu^{-1/3}) \quad \text{as } \nu \to \infty$$

and $\hat{X}_\nu \geqq X_0 > 0$ (Olver (1974, p. 438, footnote)). Now define a function $\Phi_\nu^{(j)}(z)$ by

$$\Phi_\nu^{(j)}(z) = \int_{\hat{X}_\nu}^{z} \frac{\left(t^2 - \hat{X}_\nu^2\right)^{1/2}}{t}\, dt. \tag{5.7}$$

The function $\Phi_\nu^{(j)}(z)$ has branch points at $z = 0, \pm\hat{X}_\nu$. In respect of the branch point at $z = 0$, we can either introduce a branch cut or allow $\arg z$ to take a range of values greater than 2π: we shall do both as appropriate. In respect of the branch points at $z = \pm\hat{X}_\nu$, we introduce cuts along $\operatorname{Im}\Phi_\nu^{(j)}(z) = 0$. There are thus three possible cuts emanating from each of $\pm\hat{X}_\nu$, distinguishable by the angles they make with the real axis at $\pm\hat{X}_\nu$. We define $\Phi_\nu^{(j)}(z)$ to have that cut which makes the angle $-2\pi j/3$ with the positive real axis at $\hat{X}_\nu$, and to have the corresponding cut emanating from $z = -\hat{X}_\nu$ which is obtained by reflection about the imaginary z-axis.

The branch cuts we have introduced divide the z-plane into three regions which we denote by $\mathscr{S}_\nu^{(j)}$, $j = 0, 1, 2$ (see figure 1); $j = 0$ corresponds to the eye-shaped region around the origin, while $j = 1, 2$ respectively correspond to those regions outside the eye-shaped region for which $\operatorname{Im} z > 0$ and $\operatorname{Im} z < 0$. The function $\mathscr{C}_\nu^{(j)}(z)$ is recessive in $\mathscr{S}_\nu^{(j)}$ and dominant in $\mathscr{S}_\nu^{(j-1)} \cup \mathscr{S}_\nu^{(j+1)}$; with this in mind the branch of $(t^2 - \hat{X}_\nu^2)^{1/2}$ in (5.6) is defined to be that for which $\operatorname{Im}\Phi_\nu^{(j)}(z) > 0$ in $\mathscr{S}_\nu^{(j)}$ (see (5.8)).

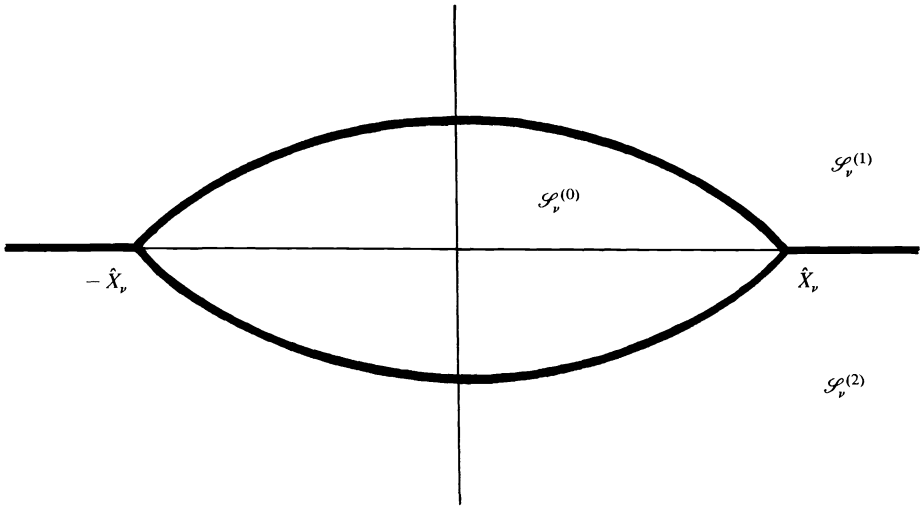


FIG. 1. *The complex z-plane. The regions $\mathscr{S}_\nu^{(j)}$ are defined to be the domains indicated, together with their boundaries.*

We define weight functions $E_\nu^{(j)}(z)$ separately in each of the three regions $\mathscr{S}_\nu^{(0)}$, $\mathscr{S}_\nu^{(1)}, \mathscr{S}_\nu^{(2)}$. In $\mathscr{S}_\nu^{(1)}$ and $\mathscr{S}_\nu^{(2)}$ we define $E_\nu^{(j)}(z)$ by

$$E_\nu^{(j)}(z) = \left|\exp\left(-i\Phi_\nu^{(j)}(z)\right)\right|, \tag{5.8}$$

where the branch of $(t^2 - \hat{X}_\nu^2)^{1/2}$ is chosen so that $E_\nu^{(j)}(z) \geqq 1$ in $\mathscr{S}_\nu^{(j)}$ and $E_\nu^{(j)}(z) \leqq 1$ in $\mathscr{S}_\nu^{(j-1)} \cup \mathscr{S}_\nu^{(j+1)}$ (cf. Olver (1974, p. 415)). In $\mathscr{S}_\nu^{(0)}$, we first define subdomains $\mathscr{D}_\nu^{(j)}(z)$ as follows. Given any point $z \in \mathscr{S}_\nu^{(0)}$, we define $\mathscr{D}_\nu^{(0)}(z)$ to be the set of points satisfying the two conditions

$$\operatorname{Im}\Phi_\nu^{(0)}(t) < \operatorname{Im}\Phi_\nu^{(0)}(z), \tag{5.9a}$$

$$-\pi < \arg t < \pi, \tag{5.9b}$$

and $\mathscr{D}_\nu^{(1)}(z)=\mathscr{D}_\nu^{(2)}(z)=\mathscr{S}_\nu^{(0)}-\mathscr{D}_\nu^{(0)}(z)$. We remark that $\Phi_\nu^{(0)}=-\Phi_\nu^{(1)}=-\Phi_\nu^{(2)}$ in $\mathscr{S}_\nu^{(0)}$. Then we define

$$e_\nu^{(j)}E_\nu^{(j)}(z)^{-1}=\sup_{\mathscr{D}_\nu^{(j)}(z)}\left\{\left|z^2-\nu^2\right|^{1/4}\left|\mathscr{C}_\nu^{(j)}(z)\right|\right\}\qquad(j=0,1,2),\tag{5.10}$$

where the constants $e_\nu^{(j)}$ are normalizing factors which are chosen to ensure the continuity of $E_\nu^{(j)}(z)$. Thus

$$e_\nu^{(j)}=\sup_{\mathscr{D}_\nu^{(j)}(\hat{X}_\nu)}\left\{\left|z^2-\nu^2\right|^{1/4}\left|\mathscr{C}_\nu^{(j)}(z)\right|\right\}\qquad(j=0,1,2).$$

(Note that since $\overline{H_\nu^{(1)}(\bar{z})}=H_\nu^{(2)}(z)$, where $\bar{\ }$ denotes complex conjugate, $e_\nu^{(1)}=e_\nu^{(2)}$.) We observe that $E_\nu^{(j)}(z)$ are continuous functions of both ν and z.

Consider further the definitions of $E_\nu^{(j)}(z)$ in $\mathscr{S}_\nu^{(0)}$. By virtue of the maximum modulus theorem we deduce that $|J_\nu(z)|$ achieves its maximum either on that part of the boundary corresponding to (5.9a) or to (5.9b). Since $|J_\nu(t)|$ is a monotonically increasing function of $|t|$ near the origin, it follows that for sufficiently small z, the maximum occurs on that part of the boundary corresponding to (5.9a). It is thus easy to deduce that $e_\nu^{(0)}E_\nu^{(0)}(z)^{-1}\sim 2|z^2-\nu^2|^{1/4}|J_\nu(z)|$ as $z\to 0$. Likewise one can show that as $z\to 0$, $e_\nu^{(1,2)}E_\nu^{(1,2)}(z)^{-1}\sim|z^2-\nu^2|^{1/4}|H_\nu^{(1)}(z)|$ or $|z^2-\nu^2|^{1/4}|H_\nu^{(2)}(z)|$.

Next we define modulus and phase functions

$$\begin{aligned}E_\nu^{(j+1)}(z)\left|\mathscr{C}_\nu^{(j+1)}(z)\right|&=M_\nu^{(j)}(z)\sin\theta_\nu^{(j)}(z),\\ E_\nu^{(j-1)}(z)\left|\mathscr{C}_\nu^{(j-1)}(z)\right|&=M_\nu^{(j)}(z)\cos\theta_\nu^{(j)}(z),\end{aligned}\tag{5.11}$$

where $M_\nu^{(j)}(z)$ is real and positive, and $\theta_\nu^{(j)}(z)$ is real. Thus

$$\begin{aligned}M_\nu^{(j)}(z)&=\left\{E_\nu^{(j+1)}(z)^2\left|\mathscr{C}_\nu^{(j+1)}(z)\right|^2+E_\nu^{(j-1)}(z)^2\left|\mathscr{C}_\nu^{(j-1)}(z)\right|^2\right\}^{1/2},\\ \theta_\nu^{(j)}(z)&=\tan^{-1}\left\{E_\nu^{(j+1)}(z)\left|\mathscr{C}_\nu^{(j+1)}(z)\right|/\left(E_\nu^{(j-1)}(z)\left|\mathscr{C}_\nu^{(j-1)}(z)\right|\right)\right\},\end{aligned}$$

where $\tan^{-1}$ takes its principal value. We remark that as $|\Phi_\nu^{(j)}(z)|\to\infty$, $\theta_\nu^{(j)}(z)\to\pi/4$ except near the boundary of $\mathscr{S}_\nu^{(j-1)}\cup\mathscr{S}_\nu^{(j+1)}$.

Likewise, we define modulus and phase functions for $\mathscr{C}_\nu^{(j)\prime}(z)$ by

$$E_\nu^{(j+1)}(z)\left|\mathscr{C}_\nu^{(j)\prime}(z)\right|=N_\nu^{(j)}(z)\sin\omega_\nu^{(j)}(z),\ E_\nu^{(j-1)}(z)\left|\mathscr{C}_\nu^{(j)\prime}(z)\right|=N_\nu^{(j)}(z)\cos\omega_\nu^{(j)}(z)\tag{5.12}$$

where $N_\nu^{(j)}(z)$ is real and positive, and $\omega_\nu^{(j)}$ is real.

It is convenient, at this stage, to introduce the constant κ which we shall require in Theorem 3. We define $\kappa=\kappa_0\kappa_1$ where

$$\begin{aligned}\kappa_0&=\sup\left\{\tfrac{1}{2}\pi\left|z^2-\nu^2\right|^{1/2}\sum_{j=0}^{2}E_\nu^{(j)}(z)^2\left|\mathscr{C}_\nu^{(j)}(z)\right|^2\right\},\\ \kappa_1&=\sup\left\{E_\nu^{(0)}(z)^{-1}E_\nu^{(1)}(z)^{-1}\right\},\end{aligned}\tag{5.13}$$

the suprema being taken over all $\nu\geqq 0$ and all $|z|>0$. The existence of κ_0 is a consequence of Lemma 3 given in Appendix B and the existence of κ_1 may be shown in a similar manner.

Now consider the functions $\mathscr{C}_{u\alpha}^{(j)}(u\zeta^{1/2})$ where the branch of $\zeta^{1/2}$ is defined in (5.5). Then under the transformation $z=u\zeta^{1/2}$, we define regions $S_\alpha^{(j)}$ in the ζ-plane corresponding to the regions $\mathscr{S}_{u\alpha}^{(j)}$ in the z-plane. These are illustrated in Fig. 2a for the case $-\frac{1}{2}\pi<\chi_0<0$.

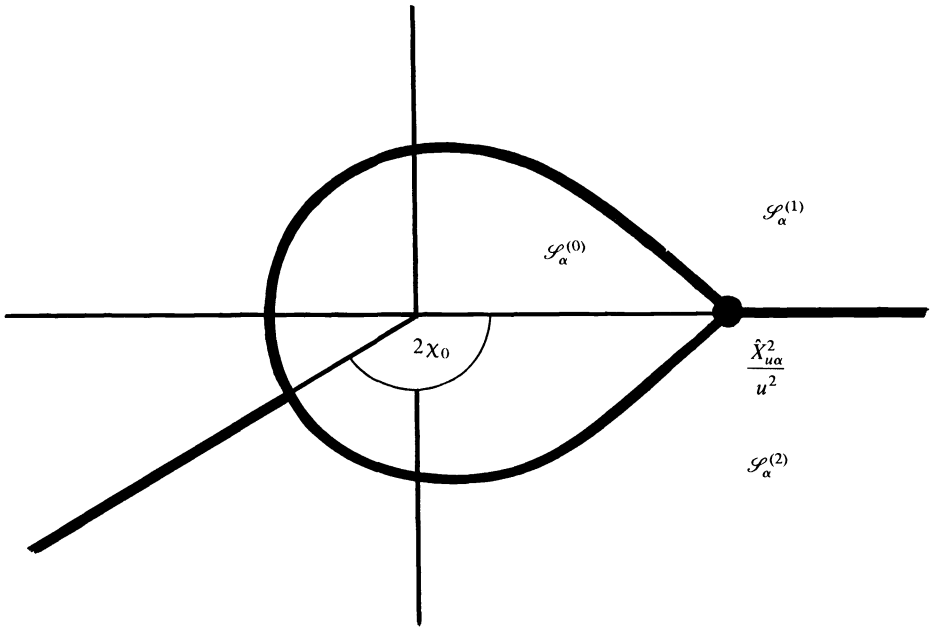


FIG. 2a. *The complex ζ-plane.*

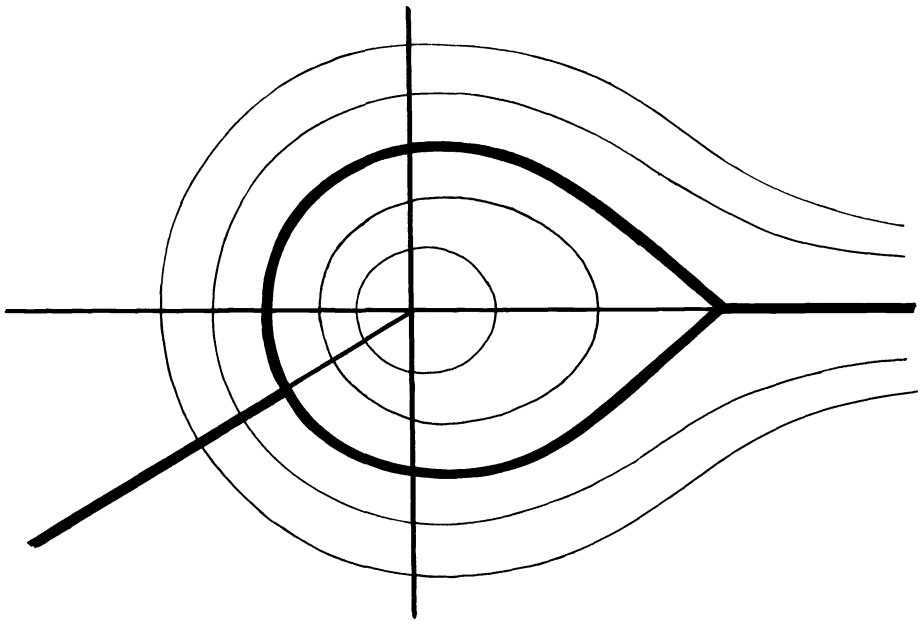


FIG. 2b. *The curves* $\operatorname{Im}\Phi_{u\alpha}^{(j)}(u\zeta^{1/2})=$ *constant.*

In order that we may state our theorem on error bounds we introduce reference points $\tilde{\zeta}^{(j)}$, $j=0,1,2$: $\tilde{\zeta}^{(0)}=0$, and $\tilde{\zeta}^{(1)}$, $\tilde{\zeta}^{(2)}$ are points (perhaps at infinity) in $S_\alpha^{(1)}$, $S_\alpha^{(2)}$ respectively, which we choose to suit our purposes. Given these $\tilde{\zeta}^{(j)}$, we define the domains $\Delta^{(j)}$, $j=0,1,2$, to be the set of points in Δ which can be linked to $\tilde{\zeta}^{(j)}$ by a path $\mathscr{P}^{(j)}$ which consists of a finite chain of arcs in R_2 having the property that as ξ passes along $\mathscr{P}^{(j)}$ from $\tilde{\zeta}^{(j)}$ to ζ,

$$\operatorname{Im}\Phi_{u\alpha}^{(j)}(u\xi^{1/2}) \geqq \operatorname{Im}\Phi_{u\alpha}^{(j)}(u\zeta^{1/2}). \tag{5.14}$$

This condition is satisfied if $\operatorname{Im}\Phi_{u\alpha}^{(j)}(u\xi^{1/2})$ is nonincreasing as ξ passes along $\mathscr{P}^{(j)}$ from $\tilde{\zeta}^{(j)}$ to ζ: this condition is useful in applications of the theory, and any such path is said to be progressive. The curves $\operatorname{Im}\Phi_{u\alpha}^{(j)}(u\xi^{1/2})=$ constant are illustrated in Fig. 2b.

THEOREM 3. *With the conditions described in* §1 *and the present section, equation* (5.3) *has, for each pair of values of* u *and* α *and each nonnegative integer* n, *solutions* $W_{2n+1}^{(j)}(u,\alpha,\zeta)$, $j=0,1,2$, *which are holomorphic in* Δ *and satisfy*

$$W_{2n+1}^{(j)}(u,\alpha,\zeta)=\zeta^{1/2}\mathscr{C}_{u\alpha}^{(j)}(u\zeta^{1/2})\sum_{s=0}^{n}\frac{A_s(\alpha,\zeta)}{u^{2s}}+\frac{\zeta}{u}\mathscr{C}_{u\alpha}^{(j)\prime}(u\zeta^{1/2})\sum_{s=0}^{n-1}\frac{B_s(\alpha,\zeta)}{u^{2s}} + \varepsilon_{2n+1}^{(j)}(u,\alpha,\zeta), \tag{5.15}$$

where

$$\frac{|\varepsilon_{2n+1}^{(j)}(u,\alpha,\zeta)|}{\zeta^{1/2}M_{u\alpha}^{(j\pm1)}(u\zeta^{1/2})},\quad \frac{|\partial\varepsilon_{2n+1}^{(j)}(u,\alpha,\zeta)/\partial\zeta|}{(1/2)uN_{u\alpha}^{(j\pm1)}(u\zeta^{1/2})+(1/2)\zeta^{-1/2}M_{u\alpha}^{(j\pm1)}(u\zeta^{1/2})} \tag{5.16}$$

$$\leqq \frac{1}{u^{2n+1}}\kappa E_{u\alpha}^{(j)}(u\zeta^{1/2})^{-1}\mathscr{V}_{\mathscr{P}^{(j)}}\left\{(\xi-\alpha^2)^{1/2}B_n(\xi)\right\}\exp\left(\frac{\kappa}{u}\mathscr{V}_{\mathscr{P}^{(j)}}\left\{(\xi-\alpha^2)^{1/2}B_0(\xi)\right\}\right),$$

when $\zeta\in\Delta^{(j)}$. *In* (5.16), *the suffix on* M *and* N *is* $j+1$ *when* $\zeta\in S_\alpha^{(j-1)}\cup S_\alpha^{(j)}$ *and* $j-1$ *when* $\zeta\in S_\alpha^{(j)}\cup S_\alpha^{(j+1)}$.

This theorem is proved in a similar manner to Theorems 1, 2. The bounds (5.16) can be sharpened somewhat by a more complicated choice of constants than the single constant defined by (5.13), but we shall not pursue this.

The bounds (5.16) can be used to deduce the asymptotic nature of the expansions (5.4); the discussion is similar to those which follow Theorems 1, 2. Thus the solution $W_{2n+1}^{(j)}(u,\alpha,\zeta)$ is recessive at $\zeta=0$ for $j=0$ and is dominant there for $j=1,2$. One can show that (5.4) is a uniformly valid compound asymptotic expansion of $W_{2n+1}^{(j)}(u,\alpha,\zeta)$. To do so, write

$$\begin{aligned} E_\nu^{(j)}(z)^{-1}M_\nu^{(j-1)}(z)&=a_\nu^{(j,-)}(z)\mathscr{C}_\nu^{(j)}(z)+b_\nu^{(j,-)}(z)z\mathscr{C}_\nu^{(j)\prime}(z),\\ E_\nu^{(j)}(z)^{-1}M_\nu^{(j+1)}(z)&=a_\nu^{(j,+)}(z)\mathscr{C}_\nu^{(j)}(z)+b_\nu^{(j,+)}(z)z\mathscr{C}_\nu^{(j)\prime}(z), \end{aligned} \tag{5.17}$$

where

$$\begin{aligned} a_\nu^{(j,-)}(z)&=E_\nu^{(j)}(z)^{-1}M_\nu^{(j-1)}(z)z\mathscr{C}_\nu^{(j+1)\prime}(z),\\ b_\nu^{(j,-)}(z)&=E_\nu^{(j)}(z)^{-1}M_\nu^{(j-1)}(z)\mathscr{C}_\nu^{(j+1)}(z),\\ a_\nu^{(j,+)}(z)&=E_\nu^{(j)}(z)^{-1}M_\nu^{(j+1)}(z)z\mathscr{C}_\nu^{(j-1)\prime}(z),\\ b_\nu^{(j,+)}(z)&=E_\nu^{(j)}(z)^{-1}M_\nu^{(j+1)}(z)\mathscr{C}_\nu^{(j-1)}(z), \end{aligned} \tag{5.18}$$

except when $j=0$: then (5.18) apply except in $\mathscr{S}_\nu^{(0)}$ where

$$a_\nu^{(0,\pm)}(z)=E_\nu^{(0)}(z)^{-1}M_\nu^{(\pm 1)}(z)/2J_\nu(z), \qquad b_\nu^{(0,\pm)}=0. \tag{5.19}$$

It then suffices to observe that for each j, $|a_\nu^{(j,-)}(z)|$ and $|b_\nu^{(j,-)}(z)|$ are uniformly bounded for $\nu\geqq 0$ in $\mathscr{S}_\nu^{(j)}\cup\mathscr{S}_\nu^{(j+1)}$, and $|a_\nu^{(j,+)}(z)|$ and $|b_\nu^{(j,-)}(z)|$ are uniformly bounded for $\nu\geqq 0$ in $\mathscr{S}_\nu^{(j-1)}\cup\mathscr{S}_\nu^{(j)}$. These results can be established by considerations similar to those motivating the proof of Lemma 3 in Appendix B.

Finally, we remark that when the domain Δ is unbounded, Theorem 3 will yield meaningful results for large ζ whenever the variations of the function $(\zeta-\alpha^2)^{1/2}B_s(\zeta)$ converge at infinity.

6. Legendre functions of large degree and order: complex variables. We use the results of §5 to construct asymptotic expansions as $\nu\to\infty$ of solutions of Legendre's equation which are uniformly valid for $0\leqq\mu/(\nu+\frac{1}{2})\leqq 1-\delta$ $(\delta>0)$ in the complex z-plane. We shall assume $\operatorname{Re} z\geqq 0$. The connection formulae for Legendre functions can be used as in §4 to derive corresponding results for other values of ν, μ, z. The pair of solutions we consider here is $P_\nu^{-\mu}(z)$ and $Q_\nu^\mu(z)$.

We proceed in a similar fashion as we did in §4: $u,\alpha,\psi(\alpha,\zeta)$ are defined as in §4, and in place of (4.2a, b) we define

$$\int_{\alpha^2}^{\zeta}\frac{(\xi-\alpha^2)^{1/2}}{2\xi}d\xi=-\int_{(1-\alpha^2)^{1/2}}^{z}\frac{(1-\alpha^2-t^2)^{1/2}}{1-t^2}dt \tag{6.1}$$

(cf. (5.1)). The functions $P_\nu^{-\mu}(z)$ and $Q_\nu^\mu(z)$ are each expressible in the form

$$\left(\frac{\zeta-\alpha^2}{1-\alpha^2-z^2}\right)^{1/4}\zeta^{-1/2}W(\zeta),$$

where $W(\zeta)$ satisfies (5.3). The domain Δ associated with (5.3) is illustrated in Fig. 3a.

A knowledge of the general configuration of the curves $\operatorname{Im}\Phi_{u\alpha}^{(j)}(u\zeta^{1/2})=\text{constant}$ in the ζ-plane is essential in applying the theory of §5. The curves, and those corresponding to them in the z-plane are shown in Figs. 3a, b (under the assumption that u is sufficiently large). In arriving at these configurations it is helpful to note that $\hat{X}_{u\alpha}/u\alpha=1+O(u^{-4/3})$ as $u\to\infty$.

It is conventional to take the branch cut associated with the Legendre functions to run along the real z-axis from $z=1$ to $z=-\infty$. With this convention, the corresponding cut in the ζ-plane is from $\zeta=0$ along the positive real axis. In the notation of §5, $\chi_0=0$ and Δ is the intersection of Δ and the ζ-plane cut along $\arg\zeta=0$.

The recessive behaviour of $P_\nu^{-\mu}(z)$ as $z\to 1$ implies that

$$P_\nu^{-\mu}(z)=c_{2n+1}^{(0)}\left(\frac{\zeta-\alpha^2}{1-\alpha^2-z^2}\right)^{1/4}\zeta^{-1/2}W_{2n+1}^{(0)}(u,\alpha,\zeta) \tag{6.2}$$

for some constant $c_{2n+1}^{(0)}$. The procedure which led to (4.11) now yields

$$c_{2n+1}^{(0)}=\frac{1}{2}e^{-\mu\pi i/2}c_{2n+1,1}. \tag{6.3}$$

It may readily be seen that the path $\mathscr{P}^{(0)}$ can be chosen to be progressive for all points in Δ (so that $\Delta^{(0)}=\Delta$). The uniform asymptotic nature of (6.2) is thus established.

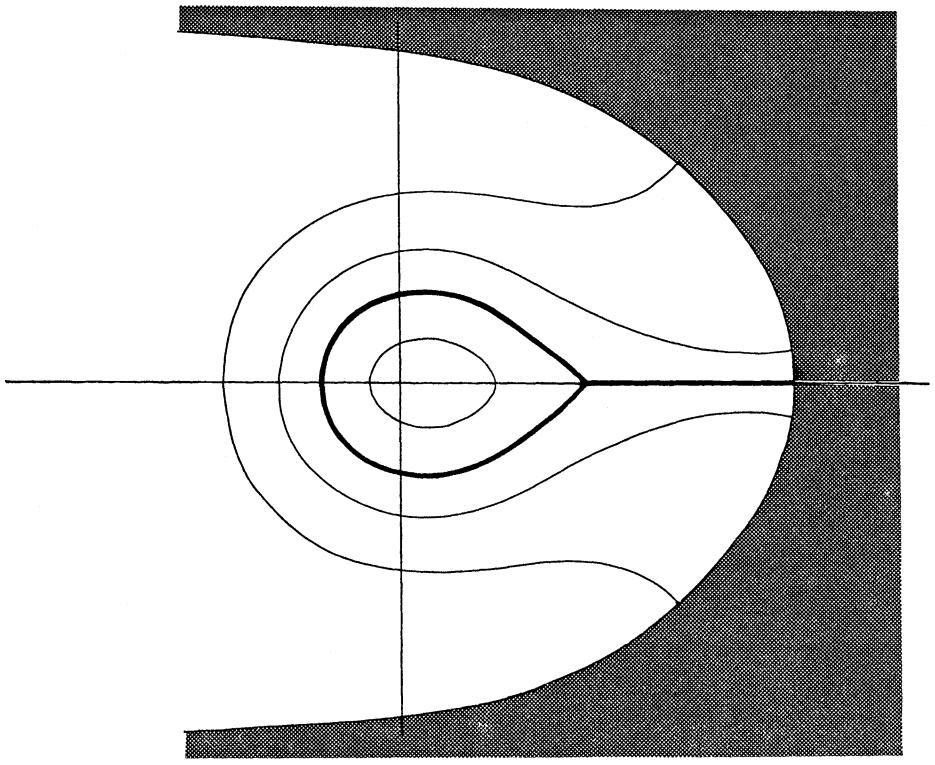


FIG. 3a. *The domain* Δ *in the complex* ζ*-plane, with curves* $\operatorname{Im} \Phi_{u\alpha}^{(j)}(u\zeta^{1/2}) = \text{constant}$.

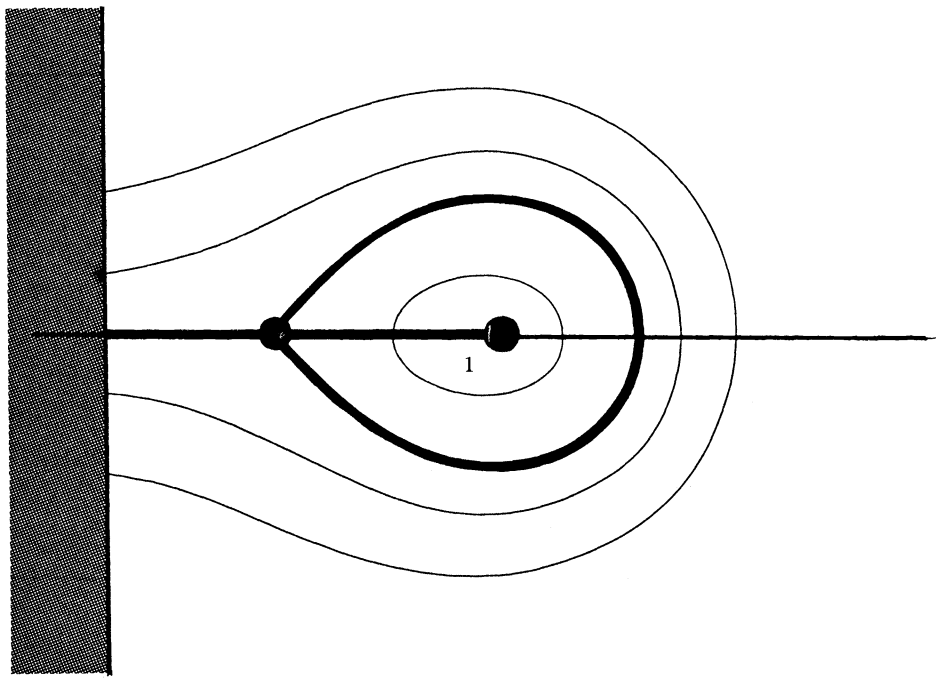


FIG. 3b. *The complex z-plane* ($\operatorname{Re} z > 0$) *with curves corresponding to* $\operatorname{Im} \Phi_{u\alpha}^{(j)}(u\zeta^{1/2}) = \text{constant}$ *in the* ζ*-plane.*

One can likewise show, by taking account of the recessive behaviour of $Q_\nu^\mu(z)$ as $z \to \infty$, that

$$Q_\nu^\mu(z) = c_{2n+1}^{(1)} \left(\frac{\zeta - \alpha^2}{1 - \alpha^2 - z^2} \right)^{1/4} \zeta^{-1/2} W_{2n+1}^{(1)}(u, \alpha, \zeta) \tag{6.4}$$

for some constant $c_{2n+1}^{(1)}$. One finds (cf. (4.18)) that

$$c_{2n+1}^{(1)} = \frac{1}{2} \pi i e^{\mu \pi i/2} c_{2n+1,4}. \tag{6.5}$$

In arriving at (6.4) the endpoint of integration is taken to be $\zeta_1 = -\infty$. It may readily be seen that the path $\mathscr{P}^{(1)}$ can be chosen to be progressive for all points in Δ (so that $\Delta^{(1)} = \Delta$). This establishes the uniform asymptotic nature of (6.4).

The results (6.2) and (6.4) can be used to determine a relation between the coefficients $c_{2n+1,1}$, $c_{2n+1,4}$, and thence the as yet undefined constants λ_s of (2.9b). If in the relation

$$i\pi \frac{\Gamma(\nu+\mu+1)}{\Gamma(\nu-\mu+1)} \mathrm{P}_\nu^{-\mu}(x) = -e^{-\mu\pi i/2} Q_\nu^\mu(x+i0) + e^{-3\mu\pi i/2} Q_\nu^\mu(x-i0), \qquad -1 < x < 1,$$

(which may be derived from the formulae of Olver (1974, Chap. (15.02), (15.03))), one substitutes the results (4.9) and (6.4), (6.5), one finds that for each n,

$$\frac{\Gamma(\nu+\mu+1)}{\Gamma(\nu-\mu+1)} c_{2n+1,1} = e^{-\mu\pi i} c_{2n+1,4}. \tag{6.6}$$

When the formulae (4.11) for $c_{2n+1,1}$ and (4.19) for $c_{2n+1,4}$ are substituted into (6.6), the result is

$$\begin{aligned} &\left(1 + \alpha \sum_{s=0}^{n-1} \frac{B_s(\alpha,0)}{(\nu+1/2)^{2s+1}} + \sum_{s=1}^{n} \frac{A_s(\alpha,0)}{(\nu+1/2)^{2s}} \right) \\ &\quad \cdot \left(1 - \alpha \sum_{s=0}^{n-1} \frac{B_s(\alpha,0)}{(\nu+1/2)^{2s+1}} + \sum_{s=1}^{n} \frac{A_s(\alpha,0)}{(\nu+1/2)^{2s}} + \delta_{2n+1,4} \right)^{-1} \\ &\quad = e^{2\mu} \left(\nu + \frac{1}{2} + \mu \right)^{-\nu-1/2-\mu} \left(\nu + \frac{1}{2} - \mu \right)^{\nu+1/2-\mu} \frac{\Gamma(\nu+\mu+1)}{\Gamma(\nu-\mu+1)}. \end{aligned} \tag{6.7}$$

An asymptotic expansion of the right-hand side in descending powers of $(\nu + \frac{1}{2})$ may be found by application of Stirling's formula; however one chooses the arbitrary constants λ_s of (2.9b), the coefficients of the corresponding power of $(\nu + \frac{1}{2})$ on the left-hand side must agree. To determine the constants λ_5, we note that the expression

$$\begin{aligned} &\left(1 + \alpha \sum_{s=0}^{n-1} \frac{B_s(\alpha,0)}{(\nu+1/2)^{2s+1}} + \sum_{s=1}^{n} \frac{A_s(\alpha,0)}{(\nu+1/2)^{2s}} \right) \\ &\quad \cdot \left(1 - \alpha \sum_{s=0}^{n-1} \frac{B_s(\alpha,0)}{(\nu+1/2)^{2s+1}} + \sum_{s=1}^{n} \frac{A_s(\alpha,0)}{(\nu+1/2)^{2s}} \right) \end{aligned}$$

is a polynomial in powers of $(\nu+\frac{1}{2})^{-2}$. We may therefore choose $\lambda_1,\lambda_2,\cdots,\lambda_n$ in turn so that the coefficients $(\nu+\frac{1}{2})^{-2},(\nu+\frac{1}{2})^{-4},\cdots(\nu+\frac{1}{2})^{-2n}$ vanish: thus

$$(6.8)\quad \left(1+\alpha\sum_{s=0}^{n-1}\frac{B_s(\alpha,0)}{(\nu+1/2)^{2s+1}}+\sum_{s=1}^{n}\frac{A_s(\alpha,0)}{(\nu+1/2)^{2s}}\right)\cdot\left(1-\alpha\sum_{s=0}^{n-1}\frac{B_s(\alpha,0)}{(\nu+1/2)^{2s+1}}+\sum_{s=1}^{n}\frac{A_s(\alpha,0)}{(\nu+1/2)^{2s}}+\delta_{2n+1,4}\right)=1+\delta_{2n+1},$$

where δ_{2n+1} is $O(u^{-2n-1})$ and may be expressed explicitly in terms of $\delta_{2n+1,4}$, $A_1,\cdots,A_n,B_1,\cdots,B_{n-1}$. Furthermore (6.7), (6.8) together imply

$$(6.9)\quad \left(1+\alpha\sum_{s=0}^{n-1}\frac{B_s(\alpha,0)}{(\nu+1/2)^{2s+1}}+\sum_{s=1}^{n}\frac{A_s(\alpha,0)}{(\nu+1/2)^{2s}}\right)(1+\delta_{2n+1})^{-1/2}$$

$$=\left(1-\alpha\sum_{s=0}^{n-1}\frac{B_s(\alpha,0)}{(\nu+1/2)^{2s+1}}+\sum_{s=1}^{n}\frac{A_s(\alpha,0)}{(\nu+1/2)^{2s}}+\delta_{2n+1,4}\right)^{-1}(1+\delta_{2n+1})^{1/2}$$

$$=e^{\mu}\left(\nu+\frac{1}{2}+\mu\right)^{-\nu/2-1/4-1/2}\left(\nu+\frac{1}{2}-\mu\right)^{\nu/2+1/4-\mu/2}\left\{\frac{\Gamma(\nu+\mu+1)}{\Gamma(\nu-\mu+1)}\right\}^{1/2}.$$

One immediately deduces from (6.9) that for our choice of λ_s, (4.11) becomes

$$(6.10)\qquad c_{2n+1,1}=\left\{\frac{\Gamma(\nu-\mu+1)}{\Gamma(\nu+\mu+1)}\right\}^{1/2}(1+\delta_{2n+1})^{-1/2}$$

and (4.19) becomes

$$(6.11)\qquad c_{2n+1,4}=\left\{\frac{\Gamma(\nu+\mu+1)}{\Gamma(\nu-\mu+1)}\right\}^{1/2}e^{\mu\pi i}(1+\delta_{2n+1})^{-1/2}.$$

Finally, we return to the problem left unsolved at the end of §4: finding a uniform asymptotic expansion for $Q_\nu^{-\mu}(x)$ in $0\leqq x<1$. Given the relation

$$Q_\nu^{-\mu}(x)=\frac{\Gamma(\nu-\mu+1)}{2\Gamma(\nu+\mu+1)}\left(e^{-\mu\pi i/2}Q_\nu^{\mu}(x+i0)+e^{-3\mu\pi i/2}Q_\nu^{\mu}(x-i0)\right)$$

[Olver (1974, Chap. 5, (13.15), (15.02))], one deduces from (6.4) that

(6.12)

$$Q_\nu^{-\mu}(x)=-\frac{1}{2}\pi c_{2n+1,1}\left(\frac{\zeta-\alpha^2}{1-\alpha^2-x^2}\right)^{1/4}\cdot\left[Y_{u\alpha}(u\zeta^{1/2})\sum_{s=0}^{n}\frac{A_s(\alpha,\zeta)}{u^{2s}}+\frac{\zeta^{1/2}}{u}Y'_{u\alpha}(u\zeta^{1/2})\sum_{s=0}^{n-1}\frac{B_s(\alpha,\zeta)}{u^{2s}}+\operatorname{Im}\left\{(\zeta+i0)^{-1/2}\varepsilon_{2n+1}^{(1)}(u,\alpha,\zeta+i0)\right\}\right],$$

and the constants λ_s are again defined implicitly by (6.8).

Appendix A. Proof of Lemma 1 (continuity of ψ). The proof is similar to that given by Olver (1975a, pp. 142–150) and to avoid undue repetition we shall simply refer to "Olver" without further specification in this appendix.

First, when the variable x is complex, the use of Cauchy's formula enables the result to be proved without difficulty (Olver, p. 142).

In the real-variable case, the proof is again relatively straightforward when $\alpha>0$ and also when $\alpha=0$, $\zeta\neq 0$. One may follow closely Olver's analysis (Olver, p. 144).

The difficult part of the proof is to establish continuity of $\psi(\alpha,\zeta)$ at the critical point $\alpha=\zeta=0$ in the real-variable case. The difficulty is essentially due to the coalescence of two critical points in the (α,ζ)-plane. In Olver's problem, these are turning points at $\zeta=\pm\alpha$. He establishes continuity by considering a deleted neighbourhood of the origin $0\leqq|\zeta|\leqq\delta$, $0\leqq\alpha\leqq\delta$, finding Taylor expansions of $x(\zeta)$ and its derivatives about each turning point, uniformly valid in α, and then appealing to his formula (2.10) for ψ. Because of the symmetry of his problem, the two proofs are effectively identical. In our problem, we proceed in a similar fashion, arriving at uniformly valid Taylor expansions of $x(\zeta)$ and its derivatives about the turning point $\zeta=\alpha^2$ and about the pole $\zeta=0$ in a deleted neighbourhood of the origin, $0\leqq|\zeta|\leqq\delta$, $0\leqq\alpha\leqq\delta^{1/2}$. More precisely we find Taylor expansions about $\zeta=\alpha^2$ for $\zeta\geqq\frac{1}{2}\alpha^2$ and about $\zeta=0$ for $\zeta\leqq\frac{1}{2}\alpha^2$. Because our problem is not symmetric with respect to the two critical points, each of the two cases has to be considered separately.

Consider first $\zeta\geqq\frac{1}{2}\alpha^2$. One finds that the proof of continuity of ψ proceeds in a very similar manner to that of Olver (pp. 145–149). This is perhaps as one would expect since in both his and our problems one is calculating Taylor expansions about a turning point.

It remains to consider $\zeta\leqq\frac{1}{2}\alpha^2$. Here there are significant differences between our analysis and Olver's. Our starting point is (2.2a) which, in terms of the function $p(\alpha,x)$ introduced in (2.6), may be written

$$\int_{\alpha^2}^{\zeta}\frac{(\alpha^2-\xi)^{1/2}}{2\xi}\,d\xi=\int_{a^2}^{x}\frac{(a^2-t)^{1/2}}{2t}\,p^{1/2}(t)\,dt, \tag{A.1}$$

where for convenience we have written $x_t=a^2$. We shall first find Taylor expansions about $\zeta=0$ in $0\leqq\zeta\leqq\frac{1}{2}\alpha^2$ and to this end we rewrite (A.1) in the form

$$\int_{\varepsilon}^{\zeta}\frac{(\alpha^2-\xi)^{1/2}}{2\xi}\,d\xi-\int_{\varepsilon/p(0)}^{x}\frac{(a^2-t)^{1/2}}{2t}\,p^{1/2}(t)\,dt = \int_{\varepsilon}^{\alpha^2}\frac{(\alpha^2-\xi)^{1/2}}{2\xi}\,d\xi-\int_{\varepsilon/p(0)}^{a^2}\frac{(a^2-t)^{1/2}}{2t}\,p^{1/2}(t)\,dt, \tag{A.2}$$

where ε is arbitrarily chosen to satisfy $0<\varepsilon<\zeta$.

The integrals on the left-hand side of (A.2) may be evaluated to yield $\frac{1}{2}\alpha\ln(\zeta/\varepsilon)+O(\zeta)$ as $\zeta\to 0$ and $-\frac{1}{2}\alpha\ln(p(0)x/\varepsilon)+O(x)$ as $x\to 0$ respectively. Since $x\to 0$ as $\zeta\to 0$, we therefore deduce that the left-hand side of (A.2) is $-\frac{1}{2}\alpha\ln(p(0)x/\zeta)$ together with a term which tends to zero as $\varepsilon\to 0$. As $\zeta\to 0$, $\varepsilon\to 0$ and the right-hand side tends to

$$\alpha\int_0^1\frac{(1-v)^{1/2}}{2v}\left\{1-\frac{p^{1/2}(a^2v)}{p^{1/2}(0)}\right\}dv.$$

We thus deduce that the ratio x/ζ is bounded as $\zeta \to 0$. The limit is actually the leading coefficient in the Taylor expansion of $x(\zeta)$. It is relatively straightforward to calculate the higher terms in the series: one finds

$$x = X_1\zeta + X_2\zeta^2 + X_3\zeta^3 + \eta, \tag{A.3}$$

where

$$X_1 = \frac{1}{p(0)} \exp \int_0^1 \frac{(1-v)^{1/2}}{v} \left\{ \frac{p^{1/2}(a^2 v)}{p^{1/2}(0)} - 1 \right\} dv,$$

$$X_2 = -\frac{1}{2\alpha^2} X_1 + \left(\frac{1}{2a^2} - \frac{1}{2} \frac{\dot{p}(0)}{p(0)} \right) X_1^2,$$

$$X_3 = \frac{1}{16\alpha^4} X_1 + \left(-\frac{1}{2} \frac{1}{a^2\alpha^2} + \frac{1}{2} \frac{p(0)}{\alpha^2 p(0)} \right) X_1^2,$$

$$+ \left(\frac{7}{16} \frac{1}{a^4} - \frac{5}{8} \frac{\dot{p}(0)}{a^2 p(0)} + \frac{7}{16} \frac{\dot{p}^2(0)}{p^2(0)} - \frac{1}{8} \frac{\ddot{p}(0)}{p(0)} \right) X_1^3.$$

In our subsequent analysis, it is necessary that these coefficients be continuous functions of α at $\alpha = 0$. To show this, we have to establish the relation between α and a. Since $x^2 f(\alpha, x) \to \alpha^2/4$ as $x \to 0$, it follows from (2.6) that

$$\alpha = a p^{1/2}(0).$$

One then readily finds that as $a \to 0$,

$$X_1 = \frac{1}{p(0)} + \frac{1}{3} \frac{\dot{p}(0)}{p^2(0)} a^2 + \left(\frac{1}{45} \frac{\dot{p}^2(0)}{p^3(0)} + \frac{1}{15} \frac{\ddot{p}(0)}{p^2(0)} \right) a^4 + O(a^6),$$

$$X_2 = -\frac{1}{3} \frac{\dot{p}(0)}{p^3(0)} + \left(-\frac{4}{15} \frac{\dot{p}^2(0)}{p^2(0)} + \frac{1}{30} \frac{\ddot{p}(0)}{p^3(0)} \right) a^2 + O(a^4), \tag{A.4}$$

$$X_3 = \frac{11}{45} \frac{\dot{p}^2(0)}{p^5(0)} - \frac{1}{10} \frac{\ddot{p}(0)}{p^4(0)} + O(a^2),$$

whence continuity follows.

We show next that $\eta = O(\zeta^4)$ in (A.3) uniformly in α. By arranging the integrals in (A.2) and taking the limit $\varepsilon \to 0$, one finds

$$I = \lim_{\varepsilon \to 0} \left\{ \int_\varepsilon^\zeta \frac{(\alpha^2 - \xi)^{1/2}}{2\xi} d\xi - J_\varepsilon \right\} - \alpha \int_0^1 \frac{(1-v)^{1/2}}{2v} \left\{ 1 - \frac{p^{1/2}(a^2 v)}{p^{1/2}(0)} \right\} dv, \tag{A.5}$$

where

$$I = \int_{X_1\zeta + X_2\zeta^2 + X_3\zeta^3}^{X_1\zeta + X_2\zeta^2 + X_3\zeta^3 + \eta} \frac{(a^2 - t)^{1/2}}{2t} p^{1/2}(t)\, dt$$

and

$$J_\varepsilon = \int_{\varepsilon/p(0)}^{X_1\zeta + X_2\zeta^2 + X_3\zeta^3} \frac{(a^2-t)^{1/2}}{2t} p^{1/2}(t)\, dt.$$

We shall show that the right-hand side of (A.5) is bounded above by $\alpha O(\zeta^3)$ and that the left-hand side is bounded below by $\alpha A|\eta/\zeta|$ for some constant A: we thus deduce $\eta = O(\zeta^4)$ uniformly.

To arrive at the first of these results, we closely follow the corresponding steps in Olver's calculation (pp. 146–147). In the integral J_ε we take a new integration variable y defined by

$$t = X_1 y + X_2 y^2 + X_3 y^3.$$

The factors in the integrand of J_ε have now to be expressed in terms of y; in particular, we find

$$a^2 - t = (\alpha^2 - y)\{Y_0 + Y_1 y + Y_2 y^2 + Y_3 y^3\} + Y_3 y^4,$$

where

$$\alpha^2 Y_0 = a^2,\ \alpha^2 Y_s = Y_{s-1} - X_s \quad \text{for } s = 1,2,3.$$

It may readily be verified that as $a \to 0$,

$$Y_0 = \frac{1}{p(0)},$$

$$Y_1 = -\frac{1}{3}\frac{\dot{p}(0)}{p^3(0)} + \left(-\frac{1}{45}\frac{\dot{p}^2(0)}{p^4(0)} - \frac{1}{15}\frac{\ddot{p}(0)}{p^3(0)}\right)a^2 + O(a^4),$$

$$Y_2 = \frac{11}{45}\frac{\dot{p}^2(0)}{p^5(0)} - \frac{1}{10}\frac{\ddot{p}(0)}{p^4(0)} + O(a^2),$$

$$Y_3 = O(1).$$

This last result enables us to claim that

$$a^2 - t = (\alpha^2 - y)\{Y_0 + Y_1 y + Y_2 y^2 + O(y^3)\}$$

uniformly for $0 < y \leqq \frac{1}{2}\alpha^2$. When all the other factors in the integrand of J_ε are expressed in powers of y we find

$$J_\varepsilon = \int_{\varepsilon_1}^{\zeta} \frac{(\alpha^2 - y)^{1/2}}{2y}\{1 + O(y^3)\}\, dy,$$

where the order term is uniformly valid in α, and ε_1 is the appropriate solution of the equation

$$\varepsilon/p(0) = X_1\varepsilon_1 + X_2\varepsilon_1^2 + X_3\varepsilon_1^3.$$

We thus deduce that the right-hand side of (A.5) is

$$\int_0^{\zeta}(\alpha^2-\xi)^{1/2}O(\xi^2)\,d\xi+\lim_{\varepsilon\to 0}\int_{\varepsilon}^{\varepsilon_1}\frac{(\alpha^2-\xi)^{1/2}}{2\xi}d\xi-\alpha\int_0^1\frac{(1-v)^{1/2}}{2v}\left\{1-\frac{p^{1/2}(a^2v)}{p^{1/2}(0)}\right\}dv,$$

and since it is easily seen that the last two terms cancel, we find that the right-hand side of (A.5) is uniformly bounded by $\alpha O(\zeta^3)$ in $0<\zeta\leqq\frac{1}{2}\alpha^2$.

Consider now the left-hand side of (A.5). By using the mean value theorem for integrals one can readily show that

$$|I|>\alpha A|\ln(1+\chi)|$$

for some constant A, uniformly in α, where

$$\chi=\eta/(X_1\zeta+X_2\zeta^2+X_3\zeta^3).$$

We showed in the previous paragraph that $I=\alpha O(\zeta^3)$ in $0<\zeta\leqq\frac{1}{2}\alpha^2$, and so it follows that in this domain

$$|\ln(1+\chi)|=O(\zeta^3)\quad\text{uniformly in }\alpha.$$

In particular we deduce that in any given neighbourhood of the origin in the (α,ζ)-plane, for $0<\zeta\leqq\frac{1}{2}\alpha^2$, the function $|\ln(1+\chi)|$ is uniformly bounded. By considering the graph of $|\ln(1+\chi)|$, we may therefore claim the existence of a constant B such that

$$|\ln(1+\chi)|>B|\chi|$$

throughout the region in question. We thus deduce that $\eta=O(\zeta^4)$ uniformly in α for $0<\zeta\leqq\frac{1}{2}\alpha^2$.

Similar arguments obtain in $\zeta<0$: one thus deduces that $\eta=O(\zeta^4)$ uniformly for $\zeta\leqq\frac{1}{2}\alpha^2$. The uniform validity of differential forms of (A.3), namely

$$\begin{aligned}x'&=X_1+2X_2\zeta+3X_3\zeta^2+O(\zeta^3),\\ x''&=2X_2+6X_3\zeta+O(\zeta^2),\\ x'''&=6X_3+O(\zeta),\end{aligned}$$

may be demonstrated, with calculations similar to those of Olver (pp. 148–149).

The continuity of ψ at the origin follows from (2.5) and the above results.

Appendix B. Existence of suprema. We shall discuss a number of results left unproved earlier. The first group of such results concern the existence of the suprema (3.5), (3.16). It suffices to consider only one of these, κ^+, in detail: the other proofs are similar. The existence of κ^+ is a consequence of the following lemma.

LEMMA 2. *Let $f(x,\nu)=\pi|x^2-\nu^2|^{1/2}M_\nu^2(x)$. Then $f(x,\nu)$ is uniformly bounded in the first quadrant of the $x\nu$-plane.*

Proof. Since $f(x,\nu)$ is continuous for $x>0$, $\nu\geqq 0$, it suffices to show

(a) $f(x,\nu)$ is uniformly bounded for $x>0$ when $\nu=0$.

(b) $\lim_{x\to 0}f(x,\nu)$ exists and is uniformly bounded for $\nu>0$.

(c) $f(x,\nu)$ is uniformly bounded as $x^2+\nu^2\to 0$ for $0<\phi<\frac{1}{2}\pi$.

(d) $f(x,\nu)$ is uniformly bounded as $x^2+\nu^2\to\infty$ for $0<\phi<\frac{1}{2}\pi$.

Here $\phi=\tan^{-1}\nu/x$.

In the following, we use several properties of $J_\nu(x)$, $Y_\nu(x)$ which can be established, for example, by perusing Olver (1974).

Condition (a).

$$f(x,0)=\begin{cases}-2\pi xJ_0(x)Y_0(x), & x\leqq X_0,\\ \pi x\big(J_0^2(x)+Y_0^2(x)\big), & x\geqq X_0.\end{cases}$$

It follows from the asymptotic behaviour of $J_0(x)$ and $Y_0(x)$ that $f(x,0)\to 0$ as $x\to 0$ and $\to 2$ as $x\to\infty$. Therefore, since $J_0(x)$ and $Y_0(x)$ are continuous for $x>0$, it follows that $f(x,0)$ is uniformly bounded for $x>0$.

Condition (b). This follows immediately from the result

$$\lim_{x\to 0} f(x,\nu)=2, \qquad \nu>0.$$

Condition (c). Define $\delta=(x^2+\nu^2)^{1/2}$. Then for $\delta<X_0$,

$$f(x,\nu)=-2\pi|x^2-\nu^2|^{1/2}J_\nu(x)Y_\nu(x).$$

We consider separately the cases $x\leqq\nu$, $x\geqq\nu$.

First, when $x\leqq\nu$ we may write

$$f(x,\nu)=-2\pi\left(1-\frac{x^2}{\nu^2}\right)^{1/2}\frac{\nu}{\sin\nu\pi}\big[\cos(\nu\pi)J_\nu^2(x)-J_\nu(x)J_{-\nu}(x)\big].$$

Now

$$J_\nu^2(x)=x^{2\nu}(1+O(\delta)), \qquad J_\nu(x)J_{-\nu}(x)=1+O(\delta^2) \quad \text{as } \delta\to 0,$$

these results being uniformly valid in x and ν, and so since $x^{2\nu}$ is uniformly bounded for $\nu>0$ and $x<1$, we deduce that $f(x,\nu)$ is uniformly bounded as $\delta\to 0$ for $x\leqq\nu$.

Next consider $x\geqq\nu$. Then we may write

$$f(x,\nu)=-2\pi\left(1-\frac{\nu^2}{x^2}\right)^{1/2}\big\{x^{1/2}J_\nu(x)\big\}\big\{x^{1/2}Y_\nu(x)\big\}.$$

It is easy to see that $x^{1/2}J_\nu(x)$ and $x^{1/2}Y_\nu(x)$ are uniformly bounded for sufficiently small x, ν. Hence $f(x,\nu)$ is uniformly bounded as $\delta\to 0$ for $x\geqq\nu$.

Actually, one can show that $f(x,\nu)\to 0$ as $\delta\to 0$ for $0\leqq\phi<\frac{1}{2}\pi$, though the limit is not uniform; as $\phi\to\frac{1}{2}\pi$ for fixed δ, $f(x,\nu)\to 2$.

Condition (d). Let us write

$$G(R,\phi)=f(x,\nu) \quad \text{when } x=R\cos\phi, \quad \nu=R\sin\phi.$$

Then

$$\text{(B.1)}\quad G(R,\phi)=\begin{cases}-2\pi R|\cos^2\phi-\sin^2\phi|^{1/2}J_{R\sin\phi}(R\cos\phi)Y_{R\sin\phi}(R\cos\phi), & \phi\geqq\phi_R,\\ \pi R|\cos^2\phi-\sin^2\phi|^{1/2}\big(J_{R\sin\phi}^2(R\cos\phi)+Y_{R\sin\phi}^2(R\cos\phi)\big), & \phi\leqq\phi_R,\end{cases}$$

where ϕ_R is the value of ϕ for which $J_{R\sin\phi}(R\cos\phi)=-Y_{R\sin\phi}(R\cos\phi)$. We wish to bound $G(R,\phi)$ as $R\to\infty$. We have the results, uniformly valid in ϕ as $R\to\infty$,

$$\text{(B.2)}\quad \begin{aligned} J_{R\sin\phi}(R\cos\phi)&\sim R^{-1/3}\left(\frac{4\zeta}{\cos^2\phi-\sin^2\phi}\right)^{1/4}\mathrm{Ai}\big(R^{2/3}\zeta\big),\\ Y_{R\sin\phi}(R\cos\phi)&\sim -R^{-1/3}\left(\frac{4\zeta}{\cos^2\phi-\sin^2\phi}\right)^{1/4}\mathrm{Bi}\big(R^{2/3}\zeta\big),\end{aligned}$$

where

$$\frac{2}{3}\zeta^{3/2} = -\int_{\sin\phi}^{\cos\phi} \frac{(\sin^2\phi - t^2)^{1/2}}{t}\,dt \qquad \left(\phi \geqq \frac{1}{4}\pi \text{ or } \zeta > 0\right),$$

$$\frac{2}{3}(-\zeta)^{3/2} = \int_{\sin\phi}^{\cos\phi} \frac{(t^2 - \sin^2\phi)^{1/2}}{t}\,dt \qquad \left(\phi \leqq \frac{1}{4}\pi \text{ or } \zeta < 0\right).$$

Let c be the negative root of the equation $\mathrm{Ai}(x) = \mathrm{Bi}(x)$ of smallest absolute value (Olver (1974, p. 395); $c = -0.366\cdots$). Effectively then, from (B.1), (B.2), $G(R,\phi)$ will be uniformly bounded as $R \to \infty$ in $0 < \phi < \frac{1}{2}\pi$ provided one can bound the following function of y:

$$4\pi|y|^{1/2}\mathrm{Ai}(y)\mathrm{Bi}(y), \qquad y \geqq c,$$
$$2\pi|y|^{1/2}[\mathrm{Ai}^2(y) + \mathrm{Bi}^2(y)], \qquad y \leqq c.$$

This is considered by Olver (1974, p. 397): the maximum occurs when $y = 1.33\cdots$ and the value of the maximum is $2.08\cdots$.

Each of conditions (a)–(d) is thus satisfied.

Next we prove the existence of κ_0, the supremum defined in (5.13). We first define, for each of $j = 0, 1, 2$,

$$f_\nu^{(j)}(z) = \frac{1}{2}\pi|z^2 - \nu^2|^{1/2} E_\nu^{(j)}(z)^2 \left|\mathscr{C}_\nu^{(j)}(z)\right|^2, \qquad z \neq 0,$$
$$f_\nu^{(j)}(0) = \frac{1}{2}\pi e_\nu^{(j)2}.$$

The existence of κ_0 is a consequence of the following lemma.

LEMMA 3. *Each of $f_\nu^{(0)}(z), f_\nu^{(1)}(z), f_\nu^{(2)}(z)$ is uniformly bounded for all $\nu \geqq 0$ and all z.*

Proof. One readily sees that $f_\nu^{(j)}(z)$ is a continuous function of ν and z for all $\nu \geqq 0$ and all z. The result will follow provided we can show that $f_\nu^{(j)}(z)$, as a function of z, is bounded throughout the z-plane including the point at infinity, and that the bound is uniform in ν.

To establish this, consider separately $z \in \mathscr{S}_\nu^{(0)}$ and $\mathscr{S}_\nu^{(1)} \cup \mathscr{S}_\nu^{(2)}$. When $z \in \mathscr{S}_\nu^{(0)}$, the definitions of $E_\nu^{(j)}(z)$ at once imply that $f_\nu^{(j)}(z) \leqq \frac{1}{2}\pi e_\nu^{(j)2}$, and consideration of the asymptotic behaviour of $\mathscr{C}_\nu^{(j)}(z)$ for large order and argument shows that $e_\nu^{(j)}$ remains bounded as $\nu \to \infty$. When $z \in \mathscr{S}_\nu^{(1)}$ or $\mathscr{S}_\nu^{(2)}$ we appeal to the inequality

$$|z^2 - \nu^2|^{1/4}\left|\mathscr{C}_\nu^{(j)}(z)\right| \leqq A\left|\exp\left(i\int_\nu^z \frac{(t^2 - \nu^2)^{1/2}}{t}\,dt\right)\right|, \tag{B.3}$$

where A is some constant independent of ν and z. Here the branch of $(t^2 - \nu^2)^{1/2}$ is defined in a similar manner as that in (5.8). The inequality (B.3) can readily be demonstrated to hold for $|\arg z| \leqq \frac{1}{2}\pi$ by using results analogous to (B.2); for $\frac{1}{2}\pi < |\arg z| \leqq \pi$ we appeal to the continuation formulae for $\mathscr{C}_\nu^{(j)}(z)$ and the result

$$\mathrm{Im}\int_{-\nu}^{\nu} \frac{(t^2 - \nu^2)^{1/2}}{t}\,dt = 0.$$

Then $f_\nu^{(j)}(z)$ will be uniformly bounded in $\mathscr{S}_\nu^{(1)}$ and $\mathscr{S}_\nu^{(2)}$ provided we can show that

$$\int_\nu^z \frac{(t^2-\nu^2)^{1/2}}{t}\,dt - \int_{\hat{X}_\nu}^z \frac{(t^2-\hat{X}_\nu^2)^{1/2}}{t}\,dt \tag{B.4}$$

is uniformly bounded for all $\nu \geqq 0$ and all $z \in \mathscr{S}_\nu^{(1)} \cup \mathscr{S}_\nu^{(2)}$. One can show, by considering separately the cases $|z/\nu|$ bounded and $|z/\nu|$ large, that (B.4) is uniformly bounded throughout $\mathscr{S}_\nu^{(1)} \cup \mathscr{S}_\nu^{(2)}$ and furthermore that the bound is $O(\nu^{-1/3})$ as $\nu \to \infty$. (Recall that $\hat{X}_\nu = \nu(1 + O(\nu^{-4/3}))$.)

The lemma follows.

Acknowledgments. We are grateful to Professor F. W. J. Olver for a number of valuable comments. This work was carried out while TMD was supported by an S.E.R.C. research studentship.

REFERENCES

[1] P. BALDWIN (1979), *Uniform approximations to solutions of a linear second order differential equation outside a vanishingly small region containing two asymptotically coincident transition points*, Quart. J. Mech. Appl. Math., 32, pp. 187–204.

[2] I. S. GRADSHTEYN AND I. M. RYZHIK (1980), *Tables of Integrals, Series and Products*, 4th edition, Academic Press, London.

[3] F. W. J. OLVER (1958), *Uniform asymptotic expansions of solutions of linear second-order equations for large values of the parameter*, Philos. Trans. Roy. Soc. London Ser. A, 250, pp. 479–517.

[4] _____ (1974), *Asymptotics and Special Functions*, Academic Press, New York.

[5] _____ (1975a), *Second-order linear differential equations with two turning points*, Phil. Trans. Roy. Soc. London Ser. A, 278, pp. 137–174.

[6] _____ (1975b), *Legendre functions with both parameters large*, Phil. Trans. Roy. Soc. London Ser. A, 278, pp. 175–185.

[7] _____ (1975c), *Unsolved problems in the asymptotic estimation of special functions*, in Theory and Application of Special Functions, R. Askey, ed., Academic Press, New York, pp. 99–141.

[8] _____ (1980), *Asymptotic expansions and error bounds*, SIAM Rev., 22, pp. 188–203.

[9] R. C. THORNE (1957a), *The asymptotic solution of linear second order differential equations in a domain containing a turning point and a regular singularity*, Phil. Trans. Roy. Soc. London Ser. A, 249, pp. 585–596.

[10] _____ (1957b), *The asymptotic expansion of Legendre functions of large degree and order*, Phil. Trans. Roy. Soc. London Ser. A, 249, pp. 597–620.

SIAM J. MATH. ANAL.
Vol. 17, No. 2, March 1986

012

ASYMPTOTIC BEHAVIOR OF THE INVARIANT DENSITY OF A DIFFUSION MARKOV PROCESS WITH SMALL DIFFUSION*

SHUENN-JYI SHEU†

Abstract. Let $X(t)$ be the n-dimensional diffusion process governed by the following equation: $dx(t)=b(x(t))\,dt+\sqrt{\varepsilon}\,dw(t)$. Assume $b(x)=\nabla I(x)+l(x)$, $\nabla I\cdot l=0$. Then under some growth conditions on I and l, we show $x(\cdot)$ has unique invariant measure with density $p^\varepsilon(x)$. And we establish the following asymptotic behavior for p^ε: $p^\varepsilon(x)=\varepsilon^{-n/2}\exp(-2I(x)/\varepsilon)(R_0(x)+O(\varepsilon))$.

Introduction. Let $x(\cdot)$ be an n-dimensional diffusion satisfying the following equation:

$$dx(t)=b(x(t))\,dt+\sqrt{\varepsilon}\,dw(t). \tag{1}$$

We will write $x^\varepsilon(\cdot)$ to indicate the dependence of $x(\cdot)$ on ε. Suppose for each $\varepsilon>0$, the diffusion process $x(\cdot)$ has invariant measure with density $p^\varepsilon(x)$. We will be interested in establishing the following limiting behavior of $p^\varepsilon(x)$:

$$\lim_{\varepsilon\to 0}\varepsilon\log p^\varepsilon(x)=-v(x), \tag{2}$$

$$p^\varepsilon(x)=\varepsilon^{-n/2}e^{-v(x)/\varepsilon}\big(T_0(x)+T_1(x)+\cdots+\varepsilon^n T_n(x)+O(\varepsilon^{n+1})\big). \tag{3}$$

A formal argument indicates that the function $v(\cdot)$ is given by the following expression:

$$v(x)=\inf_{\substack{\phi(0)=x\\ \phi(\infty)=0}}\frac{1}{2}\int_0^\infty|\dot\phi(t)+b(\phi(t))|\,dk, \tag{4}$$

where $\dot\phi(t)$ is the derivative of $\phi(t)$ at t. And T_0, T_1, $\cdots$ can be obtained by solving a family of first order partial differential equations.

This paper shows that if the function $v(x)$ given by (4) is smooth and satisfies some growth conditions, then it is possible to establish the results like (2) and a weaker result of (3), i.e.

$$\begin{aligned}&p_\varepsilon(x)=\varepsilon^{-n/2}e^{-v(x)/\varepsilon}R_\varepsilon(x),\\ &\lim R_\varepsilon(x)=R_0(x).\end{aligned} \tag{5}$$

A recent result of M. Day shows that there is a nice relation between the limiting behavior of p^ε and of the exit distribution of $x^\varepsilon(\cdot)$ from a domain (cf. [3]). A result like (5) will be useful in this respect.

* Received by the editors March 21, 1984. Part of this work was completed while the author was a visiting member at the Division of Applied Mathematics, Brown University, Providence, Rhode Island, in 1982. The work was partially supported by the National Science Foundation under contract MCS-8121940, and by the Air Force Office of Scientific Research under contract AFOSR-81-0116.

† Institute of Mathematics, Academia Sinica, Nankang, Taipei, Taiwan, Republic of China.

The plan of this paper is as follows. The assumptions are stated in §1. Under these assumptions, we give an estimation for $p^\varepsilon(\cdot)$. Section 2 shows a Ventsel–Friedlin type result (2). The basic idea is to use the relation between p^ε and the transition density. Section 3 proves the result (5). This result depends on the smoothness of the function $v(\cdot)$. However, this is a local property in the sense that if we know that (2) and the smoothness of $v(\cdot)$ on a certain region containing 0 and x, then it is possible that (5) holds.

Notation. x,y,z denote vectors in $\mathbb{R}^n$. $x\cdot y$ is the usual inner product and $|x|=(x\cdot x)^{1/2}$. For a smooth real valued function f on $\mathbb{R}^n$ $D^\alpha f(x)=(\partial^{|\alpha|}f/\partial^{\alpha_1}\cdots\partial^{\alpha_n})(x)$ for $\alpha=(\alpha_1,\cdots,\alpha_n)$, where $\alpha_i\geqq 0$ are integers.

1. Assumptions and initial results. We make the following assumptions through this paper.

Assumption (A). Assume $b(x)=-\nabla I(x)+l(x)$ where $\nabla I(x)$ is the gradient of the real valued function $I(x)$. Moreover,

(a) $\nabla I(x)\cdot l(x)=0$, $\forall x\in\mathbb{R}^n$.

(b) $I(\cdot)$, $b(\cdot)\in C^\infty$ and

$$\sup_x |D^\alpha I(x)|\leqq C_\alpha \quad \forall|\alpha|\geqq 2,$$

$$\sup_x |D^\alpha b_i(x)|\leqq C_\alpha \quad \forall|\alpha|\geqq 1,\ b(x)=(b_1(x),\cdots,b_n(x)).$$

(c) $\inf\{|\nabla I(x)|;\ |x|\geqq R\}=\delta(R)\geqq K_1 R$ if R is small, and $I(x)\to\infty$ as $|x|\to\infty$, $I(0)=0$,

(d) $b(0)=0$, $[\partial b_i(0)/\partial x_j]$ has eigenvalues with only negative real part.

LEMMA 1.1.

$$2I(x)=\inf_{\substack{\phi(0)=x\\ \phi(\infty)=0}}\frac{1}{2}\int_0^\infty |\phi(t)+b(\phi(t))|^2\,dt. \tag{1.1}$$

Proof. Since $v(x)=2I(x)$ satisfies $1/2|\nabla v(x)|^2+b(x)\cdot\nabla v(x)=0$, we realize that this is the Hamilton–Jacobi equation satisfied by the value function defined by the right-hand side of (1.1). We expect (1.1) to hold by some "verification theorem" (cf. [7]). However, it can be proved as follows.

Let $\phi^0(t)$ be the solution of

$$\frac{d\phi^0}{dt}(t)=-(b+2\nabla I)(\phi^0(t)),\qquad \phi^0(0)=x. \tag{1.2}$$

Then

$$\begin{aligned}\frac{dI(\phi^0(t))}{dt}&=-\nabla I\cdot(b+2\nabla I)(\phi^0(t))\\ &=-|\nabla I(\phi^0(t))|^2\\ &=-\frac{1}{4}|\phi^0(t)+b(\phi^0(t))|^2.\end{aligned}$$

This implies

$$I(\phi^0(T))-I(x)=-\frac{1}{4}\int_0^T |\dot{\phi}^0(t)+b(\phi^0(t))|^2\,dt,$$
$$\phi^0(T)\to 0 \quad \text{as } T\to\infty.$$

Therefore

$$I(x)=\frac{1}{4}\int_0^\infty |\dot{\phi}^0(t)+b(\phi^0(t))|^2\,dt.$$

On the other hand, for any curve $\phi(t)$,

$$\begin{aligned}|\dot{\phi}(t)+b(\phi(t))|^2&=|\dot{\phi}(t)-\nabla I(\phi(t))+l(\phi(t))|^2\\&=|\dot{\phi}(t)|^2-2\dot{\phi}(t)\cdot\nabla I(\phi(t))\\&\quad+2\dot{\phi}(t)\cdot l(\phi(t))+|\nabla I(\phi(t))|^2+|l(\phi(t))|^2\\&\geqq -4\dot{\phi}(t)\cdot\nabla I(\tau(t))\\&=-4\frac{d}{dt}I(\phi(t)).\end{aligned}$$

Therefore

$$I(x)\leqq\frac{1}{4}\int_0^\infty |\dot{\phi}(t)+b(\phi(t))|^2\,dt \quad \text{if } \phi(\infty)=0.$$

Thus we complete the proof.

Remark. For general $b(\cdot)$, without assuming (A), if the function $v(\cdot)$ defined by the right-hand side of (1.1) is smooth, then $v(\cdot)$ satisfies

$$\frac{1}{2}|\nabla v(x)|^2+b(x)\cdot V(x)=0. \tag{1.3}$$

From this we have $b(x)=-\nabla I(x)+l(x)$, where $v(x)=2I(x)$, and $\nabla I(x)\cdot l(x)=0$.

LEMMA 1.2. *Under assumption* (A), *there is a unique invariant density* $p^\varepsilon(\cdot)$ *for* $x^\varepsilon(\cdot)$. *Moreover, we have*

$$\int \exp\left(\frac{2m}{\varepsilon}I(y)\right)p^\varepsilon(y)\,dy\leqq C_m, \qquad 0<m<1 \tag{1.4}$$

for some $C_m>0$ *which is independent of* ε.

Proof. Let $f(x)=\exp((2m/\varepsilon)I(x))$ and

$$dx(t)=b(x(t))\,dt+\sqrt{\varepsilon}\,dw(t),$$
$$x(0)=x,$$
$$Lf=\frac{\varepsilon}{2}\Delta f+b\cdot\nabla f=\left(m\Delta I-\frac{2m(1-m)}{\varepsilon}|\nabla I|^2\right)f.$$

It is easy to see by our assumption that $I(x) \leqq c|x|^2$ for some $c > 0$. Then,

$$Lf(y) \leqq \begin{cases} C^{(m)} & \text{if } |y| \leqq K_m\sqrt{\varepsilon}, \\ -\alpha f(y) & \text{if } |y| \geqq K_m\sqrt{\varepsilon}. \end{cases}$$

α can be chosen as large as we want by choosing K_m such that

$$m \sup_x |\Delta I(x)| - 2m(1-m)K_1^2 K_m^2 < -\alpha.$$

Thus $f(\cdot)$ can serve as a Lyapunov function. Then $x^\varepsilon(\cdot)$ has a unique invariant density $p^\varepsilon(\cdot)$. (See [10].)

Let $\tau_R = \inf\{t: |x(t)| = R\}$. By a simple calculation

$$d(f(x(t))\exp(\alpha t)) = Lf(x(t) + \alpha f(x(t)))\exp(\alpha t)\,dt + dM(t),$$

where $M(t)$ is a martingale. Then

$$\begin{aligned} &E_x[f(x(t \wedge \tau_R))\exp(\alpha t \wedge \tau_R)] - f(x) \\ &\quad = E_x\left[\int_0^{\tau_R \wedge t} (Lf(x(s)) + \alpha f(x(s)))\exp(\alpha s)\,ds\right] \\ &\quad \leqq c\int_0^t \exp(\alpha s)\,ds \\ &\quad \leqq c_m \exp(\alpha t). \end{aligned}$$

Letting $R \to \infty$, we get

$$E_x[f(x(t))] \leqq c_m + f(x)\exp(-\alpha t).$$

We prove (1.4) by letting $t \to \infty$.

2. Ventsel–Friedlin type result. The main result of this section is the following theorem.

THEOREM 2.1.

$$\lim_{\varepsilon \to 0} \varepsilon \log p^\varepsilon(x) = -2I(x).$$

We will prove the following results:

$$\overline{\lim}\, \varepsilon \log p^\varepsilon(x) \leqq -2I(x), \tag{2.1}$$

$$\underline{\lim}\, \varepsilon \log p^\varepsilon(x) \geqq -2I(x). \tag{2.2}$$

For (2.1), we have the following estimation.

LEMMA 2.2. *There is a* $c_m > 0$ *such that*

$$\exp\left(\frac{2m}{\varepsilon} I(y)\right) p^\varepsilon(y) \leqq c_m \varepsilon^{-n/2}, \qquad 0 < m < 1.$$

Proof. We write $p^\varepsilon(t,x,y)$ and $q^\varepsilon(t,x,y)$ for the transition density of diffusion processes $x(t)$ and $y(t)$ respectively, where

$$\begin{aligned} dx(t) &= b(x(t))\,dt + \sqrt{\varepsilon}\,dw(t), \\ dy(t) &= (2\nabla I(y(t)) + b(y(t)))\,dt + \sqrt{\varepsilon}\,dw(t). \end{aligned}$$

Also, let $\pi^\varepsilon(t,x,y)=\exp((-2/\varepsilon)I(y))p^\varepsilon(t,x,y)\exp(\frac{2}{\varepsilon}I(y))$. We will obtain the following estimate:

$$p^\varepsilon(t,x,y)\leqq c_t\varepsilon^{-n/2}, \tag{2.3}$$

$$q^\varepsilon(t,x,y)\leqq c_t\varepsilon^{-n/2}, \tag{2.4}$$

$$\pi^\varepsilon(t,x,y)\leqq c_t\varepsilon^{-n/2}. \tag{2.5}$$

Assuming these, then

$$\begin{aligned} p^\varepsilon(y)&=\int p^\varepsilon(x)p^\varepsilon(t,x,y)\,dx \\ &=\int p^\varepsilon(x)\exp\left(\frac{2m}{\varepsilon}I(x)\right)\exp\left(-\frac{2m}{\varepsilon}I(x)\right)p^\varepsilon(t,x,y)\,dx. \end{aligned}$$

Together with the following inequality,

$$\begin{aligned} &\exp\left(-\frac{2m}{\varepsilon}I(x)\right)p^\varepsilon(t,x,y)\exp\left(\frac{2m}{\varepsilon}I(y)\right) \\ &\qquad=\pi^\varepsilon(t,x,y)^m p^\varepsilon(t,x,y)^{1-m}\leqq c_t\varepsilon^{-n/2}, \end{aligned}$$

we get, by using Lemma 1.2,

$$p^\varepsilon(y)\exp\left(\frac{2m}{\varepsilon}I(y)\right)\leqq c_t\varepsilon^{-n/2}\int p^\varepsilon(x)\exp\left(\frac{2m}{\varepsilon}I(x)\right)dx\leqq c_m\varepsilon^{-n/2}.$$

It remains to show (2.3), (2.4), (2.5). The proof of (2.3) and (2.4) are the same, it depends only on the estimation of the drift of the diffusion. We only consider $p^\varepsilon(t,x,y)$. It is easy to see that $p^\varepsilon(t,x,y)=\varepsilon^{-n/2}\tilde{p}^\varepsilon(t,x/\varepsilon,y/\varepsilon)$ where $\tilde{p}^\varepsilon(t,x,y)$ is the transition density of the diffusion $z(\cdot)$,

$$\begin{aligned} dz(t)&=\tilde{b}(z(t))\,dt+dw(t), \\ \tilde{b}(z)&=\frac{1}{\sqrt{\varepsilon}}b(\sqrt{\varepsilon}\,z). \end{aligned}$$

By a theorem [11, Cor. 3.37] $\tilde{p}^\varepsilon(t,x,y)\leqq c_t$. Therefore we get (2.3).

Now consider $\pi^\varepsilon(t,x,y)$. By a simple calculation, we get

$$\begin{aligned} \frac{d\pi^\varepsilon}{dt}&=\frac{\varepsilon}{2}\Delta_x\pi^\varepsilon+(2\nabla I+b)\cdot\nabla_x\pi^\varepsilon+\Delta I\cdot\pi^\varepsilon, \\ \frac{d\pi^\varepsilon}{dt}&=\frac{\varepsilon}{2}\Delta_y\pi^\varepsilon-(2\nabla I+b)\cdot\nabla_y\pi^\varepsilon-(\Delta I+\operatorname{div}b)\pi^\varepsilon, \end{aligned}$$

where ∇_x, Δ_x (resp. ∇_y, Δ_y) mean gradient and Laplacian with respect to the variable x (resp. y). From this

$$\pi^\varepsilon(t,x,y)=E_x\left[\exp\left(\int_0^t\Delta I(y(s))\,ds\right)\ y(t)=y\right]q^\varepsilon(t,x,y)$$

where $dy(t)=(2\nabla I(y(t))+b(y(t)))dt+\sqrt{\varepsilon}\,dw(t)$. Hence, by using (2.4), we obtain (2.5).

We now prove (2.2). We need a lemma from [12].

LEMMA 2.3. *Let*

$$I(T,x,y)=\inf_{\substack{\phi(0)=x\\ \phi(T)=y}}\frac{1}{2}\int_0^T|b(\phi(t))+\dot{\phi}(t)|^2\,dt.$$

Then $\lim_{\varepsilon\to 0}\varepsilon\log p^\varepsilon(T,y,x)\geqq -I(T,x,y)$ *uniformly for* x, y *on compact sets.*

Proof. See [12], or we can prove this by mimicing the arguments in [4].

LEMMA 2.4.

$$\lim_{\varepsilon\to 0}\varepsilon\log p^\varepsilon(x)\geqq -\sup_{|y|\leqq 1}I(T,x,y)\quad\forall T>0.$$

Proof. By Lemma 1.2,

$$\int_{|y|\geqq 1}p^\varepsilon(y)\,dy\to 0\quad\text{as }\varepsilon\to 0.$$

On the other hand,

$$\begin{aligned}p^\varepsilon(x)&=\int p^\varepsilon(y)p^\varepsilon(T,y,x)\,dy\\ &\geqq\int_{|y|\leqq 1}p^\varepsilon(y)p^\varepsilon(T,y,x)\,dy\\ &\geqq c(\varepsilon)\inf_{|y|\leqq 1}p^\varepsilon(T,y,x)\end{aligned}$$

where $c(\varepsilon)=\int_{|y|\leqq 1}p^\varepsilon(y)\,dy\to 1$ as $\varepsilon\to 0$. We obtain Lemma 2.4 easily by the above inequality and Lemma 2.3.

LEMMA 2.5.

$$\begin{aligned}&\lim_{T\to\infty}I(T,x,y)=v(x)=2I(x),\\ &v(x)=\inf_{\substack{\phi(0)=x\\ \phi(\infty)=0}}\frac{1}{2}\int_0^\infty|\dot{\phi}(t)+b(\phi(t))|^2\,dt.\end{aligned}$$

Proof. $v(x)=2I(x)$ was proved in Lemma 1.1. We prove the first equality as follows.

We fix x, y. Let $\phi(\cdot)$ be a curve such that $\phi(0)=x$, $\phi(T)=y$. For any $0<t<T$, following the proof of Lemma 1.1, we have

$$\begin{aligned}&\frac{1}{2}\int_0^T|\dot{\phi}(s)+b(\phi(s))|^2\,ds\geqq 2(I(\phi(t))-I(y)),\\ &\frac{1}{2}\int_0^T|\dot{\phi}(s)+b(\phi(s))|^2\,ds\geqq 2(I(x)-I(\phi(t))).\end{aligned}$$

From these we show that if there is $0<t<T$ such that $|\phi(t)|$ is large, then $\frac{1}{2}\int_0^T|\dot{\phi}(s)+b(\phi(s))|^2\,ds$ is large since $I(z)\to\infty$ as $|z|\to\infty$. And if there is $0<t<T$ such that $|\phi(t)|$ is small, then

$$\frac{1}{2}\int_0^T|\dot{\phi}(s)+b(\phi(s))|^2\,ds\geqq 2(I(x)-I(\phi(t)))\simeq 2I(x).$$

On the other hand, following an argument in [5, Lemma 3.1, p. 334], we can establish the following property:

$$\frac{1}{2}\int_0^T |\dot{\phi}+b(\phi(s))|^2 ds \to \infty \quad \text{as } T\to\infty$$

if $\delta \leqq |\phi(s)| \leqq R$, $\forall 0 \leqq s \leqq T$, where δ, R are fixed positive numbers.

By combining above results, it is easy to show

$$\varliminf_{T\to\infty} I(T,x,y) \geqq 2I(x) \quad \text{uniformly on compact sets.}$$

By using $\phi^0(\cdot)$ defined in (1.2), we can also show

$$\varlimsup_{T\to\infty} I(T,x,y) \leqq 2I(x) \quad \text{uniformly on compact sets.}$$

This completes the proof of Lemma 2.5.

3. Further results. Now consider the function R^ε defined by

$$p^\varepsilon(x) = \varepsilon^{-n/2} e^{-2I(x)/\varepsilon} R^\varepsilon(x). \tag{3.1}$$

We have the following refined result of Theorem 2.1.

THEOREM 3.1.

$$R^\varepsilon(x) \to R_0(x) \quad \textit{as } \varepsilon\to 0$$

where R_0 is the solution of the following equation:

$$-(b+2\nabla I)\cdot\nabla R_0 - (\operatorname{div} b + \Delta I) R_0 = 0, \qquad R_0(0) = c, \tag{3.2}$$

with

$$c\int \exp(-x^* A x)\, dx = 1, \qquad A = \left[\frac{\partial^2 I}{\partial x_i \partial x_j}(0)\right].$$

We compare the functions R^ε and R_0 before we give the proof of this theorem. It is not difficult to see that R^ε satisfies the following equation

$$\frac{1}{2}\varepsilon\Delta R^\varepsilon - (b+2\nabla I)\cdot\nabla R^\varepsilon - (\operatorname{div} b + \Delta I) R^\varepsilon = 0. \tag{3.3}$$

Formally, by letting $\varepsilon\to 0$ in (3.3), we get the equation (3.2) satisfied by R_0. On the other hand we have the following result which will give the condition $R_0(0)=c$.

LEMMA 3.2.

$$R^\varepsilon(\sqrt{\varepsilon}\, x) \to c \quad \textit{as } \varepsilon\to 0 \textit{ uniformly for } x \textit{ in compact sets.}$$

Proof. Let $\tilde{R}^\varepsilon(x) = R^\varepsilon(\sqrt{\varepsilon}\, x)$. Then

$$\tilde{p}(x) = \exp\bigl(-2/\varepsilon I(\sqrt{\varepsilon}\, x)\bigr)\tilde{R}^\varepsilon(x) \tag{3.4}$$

$\tilde{p}^\varepsilon$ is the invariant density for the diffusion process $y(\cdot)$

$$dy(t) = \tilde{b}^\varepsilon(y(t))\, dt + dw(t),$$

where $\tilde{b}^\varepsilon(y)=(1/\sqrt{\varepsilon})b(\sqrt{\varepsilon}y)$. Also $\tilde{p}^\varepsilon$ satisfies the following equation

$$\frac{1}{2}\Delta\tilde{p}^\varepsilon-\tilde{b}^\varepsilon\cdot\nabla\tilde{p}^\varepsilon-\operatorname{div}\tilde{b}^\varepsilon\tilde{p}^\varepsilon=0.$$

First $\tilde{p}^\varepsilon(x)\leqq c$ with c independent of ε and x by using

$$\tilde{p}^\varepsilon(x)=\int\tilde{p}^\varepsilon(y)\tilde{p}^\varepsilon(t,y,x)\,dy$$

and the fact that $\tilde{p}^\varepsilon(t,y,x)\leqq c_t$ with c_t independent of ε, x, y. Since $\tilde{b}^\varepsilon$ with all their derivatives are bounded, independent of ε, on bounded set, we can show (see [9, Theorem 6.2]) $\{\tilde{p}^\varepsilon\}$ is a compact subset of $C(\mathbb{R}^n)$ the family of all continuous real valued functions on $\mathbb{R}^n$. Also, the limit points of $\{\tilde{p}^\varepsilon\}$ are solutions of the following equation

$$\frac{1}{2}\Delta p-B\cdot\Delta p-\operatorname{div}Bp=0,$$

where $B(y)=By$ with $B=[\partial b_i/\partial x_j](0)$. On the other hand, the family of probability measures $\{\tilde{P}^\varepsilon\}$ with densities $\{\tilde{p}^\varepsilon\}$ is tight by appealing to Lemma 1.2. From these facts, $\tilde{p}$ is the only limit point of $\{\tilde{p}^\varepsilon\}$, where $\tilde{p}$ is the invariant density of

$$dy(t)=B(y(t))\,dt+dw(t)$$

and is given by

$$\tilde{p}(y)=c\exp\left(-\frac{y*Dy}{2}\right),$$

D is the only positive definite matrix satisfying $\frac{1}{2}|Dy|^2+By\cdot Dy=0$ for all $y\in R^n$.

It is not difficult to see that $D=2A$ by using the property $b(y)\cdot\nabla I(y)+|\nabla I(y)|^2=0$. Then we can get $\tilde{R}^\varepsilon(x)\to c$ by (3.4) and $\tilde{p}^\varepsilon(y)\to p(y)$ as $\varepsilon\to 0$. This completes the proof.

Proof of Theorem 3.1. Let $z(\cdot)$ be the diffusion satisfying

$$dz(t)=-(b+2\nabla I)(z(t))\,dt+\sqrt{\varepsilon}\,dw_t$$

and $\tau=\inf\{t;I(z(t))=\varepsilon R \text{ or } R\}$, where R is a fixed positive number. Then, by (3.3), we have

$$R^\varepsilon(x)=E_x\left[R^\varepsilon(z(\tau))\exp\left(-\int_0^T(\operatorname{div}b+\Delta I)(z(s))\,ds\right)\right].$$

Since $(\operatorname{div}b+\Delta I)(0)=0$, $|(\operatorname{div}b+\Delta I)(z)|\leqq c|z|$ for all z. We need the following properties:

(i) Fix R_1, $\delta_1>0$; then there is $c>0$ such that $R^\varepsilon(z)\leqq ce^{-\delta_1/\varepsilon}$ if $|z|\leqq R_1$.

(ii) There is a $\delta>0$ such that $P\{I(z(\tau))=R\}<e^{-\delta/\varepsilon}$.

(iii) $\forall p>1$, $\exists C_p>0$ with C_p independent of $\varepsilon>0$ such that

$$E_x\left[\exp\left(p\int_0^\tau|z(s)|ds\right)\right]\leqq C_p.$$

Since

$$R^\varepsilon(x)=E_x\left[R^\varepsilon(z(\tau))\exp\left(-\int_0^\tau(\operatorname{div}b+\Delta I)(z(s))\,ds\right);\ I(z(\tau))=R\right]$$
$$+E_x\left[R^\varepsilon(z(\tau))\exp\left(-\int_0^\tau(\operatorname{div}b+\Delta I)(z(s))\,ds\right);\ I(z(\tau))=\varepsilon R\right].$$

The first term on the right tends to zero because of the properties (i), (ii), (iii). On the other hand, when z is small $I(z)\simeq c_0|z|^2$ for some $c_0>0$. Therefore $R^\varepsilon(z(\tau))\to c$ as $\varepsilon\to 0$ if $I(z(\tau))=\varepsilon R$ (see Lemma 3.2). Property (iii) implies

$$\exp\left(-\int_0^\tau(\operatorname{div}b+\Delta I)(z(s))\,ds\right)\to\exp\left(-\int_0^\infty(\operatorname{div}b+\Delta I)(z_0(s))\,ds\right)$$

in L^p for any $p>1$, where $z_0(s)$ is the solution of

$$\frac{dz_0(t)}{dt}=-(b+2\nabla I)(z_0(t)).$$

Hence

$$R^\varepsilon(x)\to c\exp\left(-\int_0^\infty(\operatorname{div}b+\Delta I)(z_0(s))\,ds\right)=R_0(x).$$

It remains to show (i), (ii), (ii). (i) follows from Lemma 2.2. (ii) is a weaker result in [2]. As for (iii), we have the following lemma.

LEMMA 3.3. $E_x[\exp(p\int_0^\tau|z(s)|ds)]\leqq\exp(cpI(x)^{1/2})$ *for some* $c>0$.

Proof. Since $g(x)=E_x[\exp(p\int_0^\tau|z(s)|ds)]$ satisfies

$$Ag(x)=\frac{\varepsilon}{2}\Delta g(x)-(b+2\nabla I)\cdot\nabla g(x)+p|x|g(x)=0,$$
$$g(x)=1\quad\text{on } I(x)=\varepsilon R \text{ or } R.$$

We will choose a function $f(x)$ such that

$$\begin{aligned}&Af(x)\leqq 0\quad\text{on } \{x;\varepsilon R\leqq I(x)\leqq R\},\\ &f(x)\geqq 1\quad\text{on } I(x)=\varepsilon R \text{ or } R.\end{aligned}\tag{3.5}$$

Assuming such function $f(x)$, then, by Ito's formula

$$\begin{aligned}&d\left(f(z(\tau))\exp\left(p\int_0^t|z(s)|\,ds\right)\right)\\ &\quad=Af(z(\tau))\exp\left(p\int_0^T|z(s)|\,ds\right)dt+dM(t),\end{aligned}\tag{3.6}$$

$M(t)$ is a martingale. Equation (3.5) gives

$$E_x\left[f(z(\tau))\exp\left(p\int_0^\tau|z(s)|\,ds\right)\right]\leqq f(x)$$

by taking expectation in (3.6). Therefore

$$E_x\left[\exp\left(p\int_0^\tau|z(s)|\,ds\right)\right]\leqq f(x),$$

which gives a bound for $E_x[\exp(p\int_0^\tau|z(s)|ds)]$.

We will try $f(x)=\exp(cpI(x)^{1/2})$. A calculation gives

$$Af(x)=\left[\frac{\varepsilon}{2}\left(cp\frac{\Delta I}{I^{1/2}}(x)-cp\frac{1}{2}\frac{|\Delta I|^2}{I^{3/2}}(x)+c^2p^2\frac{|\nabla I|^2}{I}(x)\right)-cp\frac{|\nabla I|^2}{I^{1/2}}(x)+p|x|\right]f(x).$$

By assumption, on $\{x;\varepsilon R\leqq I(x)\leqq R\}$, the first term on the right is $O(\sqrt{\varepsilon})$ and

$$-cp\frac{|\nabla I|^2}{I^{1/2}}(x)+p|x|\leqq(-cp\alpha+p)|x|\leqq(-cp\alpha+p)k\sqrt{\varepsilon}$$

for some $\alpha>0$, $k>0$, where we use $I(x)>0$ if $|x|\neq 0$ and $I(x)\simeq c_0|x|^2$ if $|x|$ is small. If we choose c to be large enough, we have (3.5), then the result of the lemma.

Acknowledgment. The author would like to thank Professor W. H. Fleming for suggesting the problem and for many useful discussions.

REFERENCES

[1] M. Day (1982), *Exponential leveling for stochastically perturbed dynamical systems*, this Journal, 13, pp. 532–540.

[2] _____ (1983), *On the exponential exit law in the small parameter problem*, Stochastics, 8(4), pp. 297–323.

[3] _____, *On the asymptotic relation between equilibrium density and exit measure in the exit problem*, Stochastics, to appear.

[4] A. Friedman (1975), *Stochastic Differential Equations and Applications*, Vol. 2, Academic Press, New York.

[5] W. H. Fleming (1978), *Exit probabilities and optimal control*, Appl. Math. Optim., (4), pp. 329–346.

[6] _____ (1971), *Stochastic control for small noise intensities*, SIAM J. Control, 9, pp. 473–517.

[7] W. H. Fleming and R. W. Rishel (1975), *Deterministic and Stochastic Optimal Control*, Springer-Verlag, Berlin.

[8] C. Holland (1973), *Ergodic expansions in small noise problems*, J. Differential Equations 16, pp. 281–288.

[9] D. Gilbarg and N. S. Trudinger (1977), *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin.

[10] H. Kushner (1968), *The Cauchy problem for a class of degenerate parabolic equations and asymptotic properties of the related diffusion processes*, J. Differential Equations, 6, pp. 209–231.

[11] D. Stroock (1981), *The Malliavin calculus and its application to second order parabolic differential equations*, Math. Systems Theory, 14, pp. 25–65.

[12] S. J. Sheu, *Asymptotic expansion for the transition density of a diffusion Markov process with small diffusion*, Stochastics, to appear.

SIAM J. MATH. ANAL.
Vol. 17, No. 2, March 1986

013

ELECTROSTATICS AND THE ZEROS OF THE CLASSICAL POLYNOMIALS*

P. J. FORRESTER† AND J. B. ROGERS‡

Abstract. New interpretations of the zeros of the classical polynomials as the equilibrium positions of two-dimensional electrostatic problems are given. The electrostatic problems solved include determining the equilibrium position on the circle of $n2^M$ particles of unit charge and 2^M particles of charge $+q$, given that between every two $+q$ charges there are n unit charges.

AMS(MOS) subject classifications. Primary 82, 33

Key words. orthogonal polynomials, crystal lattices, Coulomb system

1. Introduction. Almost one hundred years ago Stieltjes gave a two-dimensional electrostatics interpretation to the zeros of the classical polynomials ([4], [5], [6]; see Szegö [7, Chap. 6]). Motivated by the desire to obtain exact information about the ground state of two-component systems interacting via the logarithmic potential, we extend Stieltjes' results. These new results supplement exact results recently obtained for the nonzero temperature statistical mechanics of such a system [1], [2].

As an illustration of our extensions, consider the unit circle in the complex plane. At the point $z=1$ fix a particle of charge $+q$ and at $z=-1$ fix a particle of charge $+p$. We show that the equilibrium position of $2N$ unit charges on the circle, symmetrically distributed about the real axis, can be written in terms of the zeros of the Jacobi polynomials. By utilizing an elementary partial fractions identity ((4.9) below), we find that the solution to more general equilibrium problems can be written down from the knowledge of the above problem. The more general problem is the determination of the equilibrium position on the circle of $n2^M$ particles of unit charge and 2^M particles of charge $+q$, given that between every two $+q$ charges there are n unit charges. We thus provide the exact structure of a local minimum of the Hamiltonian for a two-component system.

Properties of the ground states can be described in detail, as the zeros of the Jacobi polynomials have been extensively studied [7]. Conversely, many intuitively obvious properties of the electrostatic problem can be formulated as theorems regarding the zeros of the Jacobi polynomials, thus providing electrostatic interpretations to those theorems.

2. Two impurities on the circle. Consider a system of $2N$ particles of charge $+1$, labelled $\theta_1, \theta_2, \cdots, \theta_{2N}$, confined to a circle of radius R interacting via the potential

$$V(\theta_j,\theta_k) = -q_k \log R|e^{i\theta_j}-e^{i\theta_k}|. \tag{2.1}$$

At $\theta=0$ fix a particle of charge $+q$ and at $\theta=\pi$ fix a particle of charge $+p$. The uniform neutralizing background necessary to obtain thermodynamic stability is not

* Received by the editors February 28, 1984.

† Department of Theoretical Physics, Research School of Physical Sciences, The Australian National University, Canberra, Australia 2601.

‡ Department of Mathematics, University of Melbourne, Parkville, Victoria, Australia 3052.

relevant when calculating the ground state equilibrium positions, since it only contributes a constant to the Hamiltonian. Thus ignoring constant terms, we may take for the Hamiltonian

$$H_0 = -q\sum_{k=1}^{2N}\log|1-e^{i\theta_k}| - p\sum_{k=1}^{2N}\log|1+e^{i\theta_k}| - \sum_{1\leq k<j\leq N}\log|e^{i\theta_k}-e^{i\theta_j}|. \tag{2.2}$$

Note that the radius has been scaled out of the Hamiltonian. We seek the minimum of H_0 subject to the requirement

$$0<\theta_j<\pi, \quad j=1,2,\cdots,N \quad \text{and} \quad \pi<\theta_j<2\pi, \quad j=N+1,\cdots,2N. \tag{2.3}$$

Such a minimum exists and is unique [7, p. 140]. By symmetry $\theta_j=-\theta_{2N+1-j}$, $j=1,2,\cdots,N$. The location of the minimum is given by the following result.

THEOREM 2.1. *The minimum of the function H_0 subject to the constraint* (2.3) *occurs at the zeros of the Jacobi polynomial $P_N^{(q-1/2,p-1/2)}(\cos\theta)$, $0<\theta<2\pi$.*

Proof. Using the identity

$$|e^{i\theta_j}-e^{i\theta_k}| = i^{-1}(e^{i\theta_j}-e^{i\theta_k})\exp\left(-\frac{i}{2}(\theta_j+\theta_k)\right), \qquad \theta_j\geq\theta_k, \tag{2.4}$$

we see that the condition for a minimum is the set of nonlinear equations

$$0=\frac{\partial H}{\partial\theta_k} = -i\Bigg(qe^{i\theta_k}/(e^{i\theta_k}-1)+pe^{i\theta_k}/(e^{i\theta_k}+1) - (2N-1+p+q)/2+\sum_{j=1}^{2N}{}'\, e^{i\theta_k}/(e^{i\theta_k}-e^{i\theta_j})\Bigg), \qquad 1\leq k\leq 2N, \tag{2.5}$$

where the prime on the sum indicates that the $j=k$ term is to be omitted. Following Stieltjes, consider the polynomial

$$f(x)=\prod_{l=1}^{2N}(x-e^{i\theta_l}), \tag{2.6}$$

which has its zeros at the minimum of H_0. A short calculation then shows

$$\frac{f''(e^{i\theta_k})}{f'(e^{i\theta_k})} = 2\sum_{j=1}^{N}{}'\,\frac{1}{(e^{i\theta_k}-e^{i\theta_j})}. \tag{2.7}$$

Thus we can write the ground state condition (2.5) as

(2.8)

$$0=e^{i\theta_k}(e^{2i\theta_k}-1)f''(e^{i\theta_k}) + \left(2qe^{i\theta_k}(e^{i\theta_k}+1)+2pe^{i\theta_k}(e^{i\theta_k}-1)-(2N-1+p+q)(e^{2i\theta_k}-1)\right)f'(e^{i\theta_k}), \qquad 1\leq k\leq 2N.$$

The $(2N+1)$th order polynomial

$$x(x^2-1)f''(x)+\left(2qx(x+1)+2px(x-1)-(2N-1+p+q)(x^2-1)\right)f'(x) \tag{2.9}$$

therefore has zeros at $x=e^{i\theta_k}$, $k=1,2,\cdots,2N$, and so is proportional to $(x-A)f(x)$ where A is a constant. By equating like powers of x^{2N+1} we find the proportionality constant to be $2N(p+q)$, so

$$(2.10)\quad x(x^2-1)f''(x)+\big(2qx(x+1)+2px(x-1)-(2N-1+p+q)(x^2-1)\big)f'(x) \\ -2N(p+q)(x-A)f(x)=0.$$

It follows from the uniqueness of a minimum of H_0 that there is one and only one value of A such that (2.6) satisfies (2.10) with the θ_k constrained by (2.3). When $N=1$ it is a simple exercise to derive

$$(2.11)\qquad A=(p-q)/(p+q).$$

We now verify that the choice (2.11) gives the required polynomial solution of (2.10) for general N. By making standard transformations applicable to second-order differential equations [7, p. 16] and then changing variables $x=e^{i\theta}$, we write (2.10) with A given by (2.11) in the form

$$(2.12)\quad u''+\big(q(1-q)/(4\sin^2\theta/2)+p(1-p)/(4\cos^2\theta/2)+(2N+p+q)^2/4\big)u=0,$$

where

$$(2.13)\qquad u(\theta)=(\sin^q\theta/2)(\cos^p\theta/2)e^{-iN\theta}f(e^{i\theta}).$$

However (2.12) is also a transformed version of the differential equation of the Jacobi polynomials [7, p. 67] and is satisfied by

$$(2.14)\qquad u(\theta)=(\sin^q\theta/2)(\cos^p\theta/2)\,P_N^{(q-1/2,p-1/2)}(\cos\theta).$$

Hence we have

$$(2.15)\qquad f(e^{i\theta})=e^{iN\theta}P_N^{(q-1/2,p-1/2)}(\cos\theta),$$

which is the desired polynomial solution of (2.10) satisfying the constraint (2.3).

Theorem 2.1 is to be compared to Stieltjes' interpretation [7, p. 140].

THEOREM 2.2 (Stieltjes). *The minimum of the function*

$$(2.16)\qquad T=-\sum_{k=1}^{N}\big(\log(1+x_k)^P+\log(1-x_k)^Q\big)-\sum_{1\le k<j\le N}\log|x_k-x_j|,$$

$P, Q>0$, $-1<x_k<1$, *occurs at the zeros of the Jacobi polynomial* $P_N^{(2Q-1,2P-1)}(x)$.

The expression T is the Hamiltonian for N unit charges confined to the interval $[-1,1]$ by a particle of charge P fixed at $x=1$, and a particle of charge Q fixed at $x=-1$, interacting via the logarithmic potential. Thus comparing Theorems 2.1 and 2.2, we see that the equilibrium positions are closely related. Consider the unit circle in the complex plane. At $z=1$ fix a particle of charge $q=2Q-\frac{1}{2}$, and at $z=-1$ a particle of charge $p=2P-\frac{1}{2}$ ($p,q\ge 0$). If the equilibrium positions of the $2N$ charges as given by Theorem 2.1 are projected orthogonally onto the real axis, then we locate the equilibrium positions of Stieltjes' problem, Theorem 2.2.

As an example of a theorem regarding the zeros of the Jacobi polynomials which is intuitively obvious when viewed from the electrostatic interpretation, consider the following result [7, p. 121].

THEOREM 2.3 (Markoff, Stieltjes). *Let* $\{x_\nu = x_\nu(\alpha,\beta)\}$ *denote the zeros of the Jacobi polynomial* $P_N^{(\alpha,\beta)}(x)$. *Then*

$$\text{(i)} \quad \frac{\partial x_\nu}{\partial \alpha} < 0, \qquad \text{(ii)} \quad \frac{\partial x_\nu}{\partial \beta} > 0.$$

From Stieltjes' interpretation of the zeros of $P_N^{(\alpha,\beta)}$, this is equivalent to saying that (i) when the charge at $x=1$ is increased, the N unit charges are repelled towards the fixed charge at $x=-1$; (ii) when the charge at $x=-1$ is increased, the N unit charges are repelled towards the fixed charge at $x=1$. Thus Theorem 2.3 is intuitively obvious.

Reconsider Theorem 2.1. The positions of the unit charges around an impurity (the $+q$ charge at $\theta=0$, say) in the large N limit are of particular interest. We have the following theorem [7, p. 193].

THEOREM 2.4. *Let* $x_{1N} > x_{2N} > \cdots$ *be the zeros of* $P_N^{(\alpha,\beta)}(x)$ *in* $[-1,+1]$ *in decreasing order* (α,β *real but otherwise arbitrary*). *If we write* $x_{\nu N} = \cos\theta_{\nu N}$, $0 < \theta_{\nu N} < \pi$, *then for fixed* ν,

$$\lim_{N\to\infty} N\theta_{\nu N} = j_{\nu,\alpha},$$

where $j_{\nu,\alpha}$ *denotes the* ν*th positive zero of* $J_\alpha(z)$, J_α *denoting the Bessel function of order* α.

Regarding the $N\to\infty$ limit as the thermodynamic limit, with particle density π/N, we have from Theorems 2.1 and 2.4 the following result.

THEOREM 2.5. *Consider the real line with a particle of charge* $+q$ *fixed at the origin, and particles of unit charge at unit density immersed in a uniform neutralizing background. In the thermodynamic limit the ground state positions of the unit charges are given by*

$$\pm j_{\nu,q-1/2}/\pi, \qquad \nu = 1,2,\cdots.$$

3. A single impurity on the line. In §2 it was shown how an equilibrium problem on the circle gave rise to the Jacobi polynomials. In this section we consider an analogous physical problem, in which the equilibrium position of the charges are given in terms of the Laguerre polynomials.

Consider $2N$ particles of unit charge free to move on the real line with a particle of charge $+q$ fixed at the origin. The system is made electrically neutral by imposing a background charge density of profile

$$\sigma(y) = \begin{cases} -(c/\pi)(1-y^2/L^2)^{1/2}, & |y|<L, \\ 0, & |y|\geqq L, \end{cases} \tag{3.1}$$

$c = 2(2N+q)/L$.

To compute the Hamiltonian for the system we require the integral [3, p. 44]

$$\frac{1}{\pi}\int_{-L}^{L} dy(1-y^2/L^2)^{1/2}\log|x-y| = x^2/2L + K, \tag{3.2}$$

where K is independent of x. This gives us the contribution to the Hamiltonian of the particle-background interaction. Thus if we ignore constant terms, the Hamiltonian is given by

$$H_1=\left(\frac{c}{2L}\right)\sum_{j=1}^{2N}x_j^2-q\sum_{k=1}^{2N}\log|x_k|-\sum_{1\leqq j<k\leqq 2N}\log|x_j-x_k|, \tag{3.3}$$

where we have labelled the $2N$ unit charges $x_1, x_2, \cdots, x_{2N}$.

We seek the minimum of H_1 subject to the condition that N particles are on either side of the impurity at the origin; that is,

$$x_k<0, \quad k=1,2,\cdots,N \quad \text{and} \quad x_k>0, \quad k=N+1,\cdots,2N. \tag{3.4}$$

It is a simple exercise to prove that such a minimum exists and is unique [7, p. 140]. By symmetry we must have

$$-x_k=x_{N+k}, \qquad k=1,2,\cdots,N. \tag{3.5}$$

The location of the minimum is given by the following result.

THEOREM 3.1. *The minimum of the function H_1 subject to the constraint* (3.5) *occurs when the x_k are the zeros of the Laguerre polynomial*

$$L_N^{(q-1/2)}(cx^2/L). \tag{3.6}$$

Proof. The condition for a minimum is the set of nonlinear equations

$$0=\frac{\partial H_1}{\partial x_k}=(c/L)x_k-\sum_{j=1}^{2N}{}'\frac{1}{x_k-x_j}-\frac{q}{x_k}, \qquad k=1,\cdots,2N. \tag{3.7}$$

Consider the polynomial

$$g(x)=\prod_{l=1}^{2N}(x-x_l) \tag{3.8}$$

which has its zeros at the minimum of H_1. Then we can write (3.7) as

$$0=x_kg''(x_k)+\left(-(2c/L)x_k^2+2q\right)g'(x_k), \qquad k=1,\cdots,2N. \tag{3.9}$$

Thus we have a $(2N+1)$th order polynomial which has the $2N$ zeros of $g(x)$. Hence

$$xg''(x)+\left(-(2c/L)x^2+2q\right)g'(x)+a(x-b)g(x)=0. \tag{3.10}$$

By equating like coefficients of x^{2N}, we find $a=4Nc/L$. Furthermore the symmetry condition (3.5) requires g to be even. Using this condition in (3.10) gives $b=0$.

Replacing x by $(Lx/c)^{1/2}$ the differential equation assumes the form

$$xu''+\left(-x+(q+1/2)\right)u'+Nu=0, \tag{3.11}$$

$$u(x)=g\left((Lx/c)^{1/2}\right). \tag{3.12}$$

We recognize (3.11) as the differential equation satisfied by the Laguerre polynomial [7, p. 100] so that

$$u(x)=L_N^{(q-1/2)}(x). \tag{3.13}$$

If we substitute (3.13) in (3.12), Theorem 3.1 follows immediately.

As $L\to\infty$ and y remains finite, $\sigma(y)\to -c/\pi$. Therefore in the thermodynamic limit if we choose $c=\pi$ (and thus $L=2(2N+q)/\pi^2$), the background around the origin tends to a constant unit density. From Theorem 2.4 we know the equilibrium position of unit charges immersed in a neutralizing background at unit density around a $+q$ impurity on the line. Hence from the equivalence of the two physical problems in the thermodynamic limit, we can deduce the well-known theorem relating the zeros of the Laguerre polynomials to the zeros of the Bessel functions [7, p. 193].

THEOREM 3.2. *Let $y_{1N}<y_{2N}<\cdots$ be the zeros of $L_N^{(q-1/2)}(y)$ in increasing order. Then*

$$\lim_{N\to\infty} Ny_{lN}=\frac{1}{4}\left(j_{l,q-1/2}\right)^2.$$

4. Crystal lattices on the circle. In this section crystal lattice structures are given for the equilibrium positions on the circle of two-component systems. By a crystal lattice we mean a lattice which has a cell repeated periodically. We have the following result.

THEOREM 4.1. *Suppose there are $n2^M$ particles of unit charge and 2^M particles of charge $+q$ on the circle with one of the $+q$ charges fixed at $\theta=0$. Require that between every two $+q$ charges there are n unit charges. Then the equilibrium position of the $n2^M$ particles of unit charge are given by the zeros of*

$$P_{n/2}^{((q-1)/2,-1/2)}(\cos 2^M\theta), \qquad 0<\theta<2\pi \tag{4.1}$$

if n is even, and the zeros of

$$P_{(n-1)/2}^{((q-1)/2,1/2)}(\cos 2^M\theta), \qquad 0<\theta<2\pi \tag{4.2}$$

if n is odd. The equilibrium position of the (2^M-1) particles of charge $+q$ occurs at

$$\theta=\frac{2\pi k}{2^M}, \qquad k=1,2,\cdots,(2^M-1). \tag{4.3}$$

Proof. Label the coordinates of the particles of charge $+q$ $\phi_0=0$, $\phi_1,\cdots,\phi_{2^M-1}$ and the particles of unit charge θ_1, $\theta_2,\cdots,\theta_{n2^M}$. Then the Hamiltonian for the system is

$$\begin{aligned} H_2=&-q^2\sum_{0\le j<k\le 2^M-1}\log|e^{i\phi_j}-e^{i\phi_k}| \\ &-q\sum_{0\le j\le 2^M-1}\ \sum_{1\le k\le n2^M}\log|e^{i\phi_j}-e^{i\phi_k}|-\sum_{1\le j<k\le 2^M}|\log e^{i\theta_j}-e^{i\theta_k}|. \end{aligned} \tag{4.4}$$

Again it is easy to show the existence and uniqueness of the required minimum [7, p. 140]. To prove the theorem we must show that the equations for a minimum, obtained by taking partial derivatives of H_2, are satisfied when the angles are given by

(4.1), (4.2) and (4.3). Explicitly we must show

$$0 = q \sum_{j=0}^{2^M-1} \left(e^{i(\theta_{k'}+2\pi l')/2^M} / e^{i(\theta_{k'}+2\pi l')/2^M} e^{2\pi i j/2^M} \right) \tag{4.5}$$

$$+ \sum_{k=1}^{n} \sum_{l=0}^{2^M-1}{}' e^{i(\theta_{k'}+2\pi l')/2^M} / \left(e^{i(\theta_{k'}+2\pi l')/2^M} - e^{i(\theta_k+2\pi i l)/2^M} \right)$$

$$- \left[(n+q)2^M - 1 \right]/2, \qquad 1 \leqq k' \leqq n, \quad 0 \leqq l' \leqq 2^M - 1,$$

and

$$0 = q \sum_{k=1}^{n} \sum_{l=0}^{2^M-1} e^{2\pi i k'/2^M} / \left(e^{(2\pi i k'/2^M)} - e^{i(\theta_k+2\pi l)/2^M} \right) \tag{4.6}$$

$$+ q^2 \sum_{\substack{k=0 \\ k \neq k'}}^{2^M-1} e^{2\pi i k'/2^M} / \left(e^{2\pi i k'/2^M} - e^{2\pi i k/2^M} \right)$$

$$- \left(q^2(2^M-1) + qn2^M \right)/2, \qquad 0 \leqq k' \leqq 2^M - 1,$$

where here the θ_k are the zeros of

$$P_{n/2}^{((q-1)/2,-1/2)}(\cos\theta), \qquad 0 < \theta < 2\pi, \tag{4.7}$$

if n is even, and the zeros of

$$P_{(n-1)/2}^{((q-1)/2,1/2)}(\cos\theta), \qquad 0 < \theta < 2\pi, \tag{4.8}$$

if n is odd.

We will reduce (4.5) to the solution of the impurity problem, Theorem 2.1. Consider the identity

$$\frac{e^{i\theta}}{e^{i\theta} - e^{i\phi}} = \frac{1}{2} \left(\frac{e^{i\theta/2}}{e^{i\theta/2} - e^{i\phi/2}} + \frac{e^{i\theta/2}}{e^{i\theta/2} - e^{i(\phi+2\pi)/2}} \right). \tag{4.9}$$

Applying (4.9) iteratively M times we have the identity

$$e^{i\theta} / \left(e^{i\theta} - e^{i\phi} \right) = \left(\frac{1}{2^M} \right) \sum_{j=0}^{2^M-1} \frac{e^{i\theta/2^M}}{\left(e^{i\theta/2^M} - e^{i(\phi+2\pi j)/2^M} \right)}. \tag{4.10}$$

Substitution of (4.10) in the first sum of (4.5) with $\theta = \theta_{k'} + 2\pi i l'$ and $\phi = 0$, and the second sum of (4.5) with $\theta = \theta_{k'} + 2\pi i l'$ and $\phi = \theta_k + 2\pi i l$ $(k \neq k')$ reduces (4.5) to

$$0 = \frac{q e^{i\theta_{k'}}}{(e^{i\theta_{k'}} - 1)} - \frac{\left((n+q)2^M - 1 \right)}{2^{M+1}} \tag{4.11}$$

$$+ \sum_{k=1}^{n} \frac{e^{i\theta_{k'}}}{\left(e^{i\theta_{k'}} - e^{i\theta_k} \right)} + 2^{-M} \sum_{l=0}^{2^M-1}{}' \frac{e^{2\pi i l'/2^M}}{\left(e^{2\pi i l'/2^M} - e^{2\pi i l/2^M} \right)}.$$

However, it is a simple exercise in summing series to prove that the last sum equals $(2^M-1)/2$, so we reclaim (2.5) (with $p=0$ and $n=2N$ if n is even, $p=1$ and $n-1=2N$ if n is odd), which we know from Theorem 2.1 is satisfied by the choices (4.7) and (4.8).

To prove (2.6), note that the second sum is equal to $(2^M-1)/2$ and that the first sum can be reduced using the identity (4.10) with $\theta=2\pi k'$ and $\phi=\theta_k+2\pi l$. Thus we are required to show

$$0=-n/2+\sum_{k=1}^{n}1/(1-e^{i\theta_k}). \tag{4.12}$$

This follows immediately from the fact that the θ_k are situated symmetrically about the real axis, and from the identity

$$1/(1-e^{i\theta_k})+1/(1-e^{-i\theta_k})=1. \tag{4.13}$$

If we regard the $N\to\infty$ limit as the thermodynamic limit with particle density of the $+q$ species $\pi/2^{M-1}$, we have from Theorem 4.1 the following result.

THEOREM 4.2. *Consider the real line with n particles of unit charge arranged between every pair of particles of charge $+q$. Suppose the particles of charge $+q$ are at unit density and all the charges are immersed in a neutralizing uniform background. Then in the thermodynamic limit the ground state for the system is a crystal lattice with cells of unit length. When n is even, a cell is specified by a $+q$ charge at $x=0$, with the n unit charges at the zeros of $P_{n/2}^{((q-1)/2,\,-1/2)}(\cos 2\pi x)$, $0<x<1$, while when n is odd a cell is specified by a $+q$ charge at $x=0$, with the n unit charges at the zeros of $P_{(n-1)/2}^{((q-1)/2,\,1/2)}(\cos 2\pi x)$, $0<x<1$.*

REFERENCES

[1] P. J. FORRESTER, *An exactly solvable two component classical Coulomb system*, J. Austral. Math. Soc., Ser. B, 26(1984), pp. 119–128.

[2] _____, *Interpretation of an exactly solvable two component plasma*, J. Stat. Phys., 35 (1984), pp. 77–87.

[3] M. L. MEHTA, *Random Matrices*, Academic Press, New York, 1967.

[4] T. J. STIELTJES, *Sur quelques théorèmes d'algèbre*, Oeuvres Complètes, Vol. 1, pp. 440–441.

[5] _____, *Sur les polynômes de Jacobi*, Oeuvres Complètes, Vol. 1, pp. 442–444.

[6] _____, *Sur les racines de l'equation $X_n=0$*, Oeuvres Complètes, Vol. 2, pp. 73–88.

[7] G. SZEGÖ, *Orthogonal Polynomials*, 3rd ed., American Mathematical Society, Providence, RI, 1967.

SIAM J. MATH. ANAL.
Vol. 17, No. 2, March 1986

014

ON THE ORDER OF MAGNITUDE OF FOURIER COEFFICIENTS*

J. B. READE†

Abstract. The main result is that the Fourier coefficients of any continuous function are $o(1/n)$ if a monotonicity condition is assumed. The proof is achieved by showing the coefficients are always $o(1/n)$ in a certain weaker sense, and then proving a Tauberian theorem giving $o(1/n)$ in the ordinary sense under the extra assumption of monotonicity.

1. Let $f(t)\in L^1[-1,1]$ have Fourier cosine coefficients

$$a_n=\int_{-1}^{1} f(t)\cos n\pi t\,dt.$$

The Riemann–Lebesgue lemma says that $a_n\to 0$ as $n\to\infty$. The convergence can be arbitrarily slow, since, for any sequence (k_n) such that $|k_n|\to\infty$ as $n\to\infty$, we have

$$k_n a_n=\int_{-1}^{1} f(t)k_n\cos n\pi t\,dt=T_n f,$$

where T_n is a linear functional on $L^1[-1,1]$ having norm

$$\|T_n\|=\max_{|t|\leqq 1}|k_n\cos n\pi t|=|k_n|,$$

so $k_n a_n\to 0$ for every $f\in L^1[-1,1]$ would contradict the uniform boundedness principle. The same is true for $f(t)\in C[-1,1]$ for similar reasons.

It turns out, however, that one can make positive assertions about the order of magnitude of Fourier cosine coefficients if one interprets, e.g., $na_n\to 0$ in a weaker sense. Specifically, we shall show firstly that, if $f(t)\in L^2[-1,1]$, then $n^{1/2}a_n\to 0$ in the sense of Cesaro, i.e.,

$$\frac{a_1+2^{1/2}a_2+\cdots+n^{1/2}a_n}{n}\to 0,$$

and secondly that, if $f(t)\in C[-1,1]$, then $na_n\to 0$ in the double Cesaro sense, i.e., if

$$u_n=\frac{a_1+2a_2+\cdots+na_n}{n},$$

$$v_n=\frac{u_1+u_2+\cdots+u_n}{n},$$

then $v_n\to 0$. It is possible to deduce, e.g., that, if $a_n\geqq a_{n+1}\geqq 0$ for all $n\geqq 0$, then convergence is in the ordinary sense in both cases.

We also prove similar results for Fourier sine coefficients

$$b_n=\int_{-1}^{1} f(t)\sin n\pi t\,dt,$$

though in some cases the proofs are different.

*Received by the editors December 12, 1983, and in revised form June 15, 1984.
†Department of Mathematics, The University, Manchester M13 9PL, England.

2. The result concerning $f(t)\in L^2[-1,1]$ depends on the Riesz–Fischer theorem and two lemmas about Cesaro convergence.

LEMMA 1. *For any convergent series* $\Sigma_1^\infty a_n$, *we have* $na_n\to 0$ *in the sense of Cesaro.*

Proof. Let $s_n=a_1+a_2+\cdots+a_n\to s$. Then

$$\frac{a_1+2a_2+\cdots+na_n}{n}=\frac{s_1+2(s_2-s_1)+\cdots+n(s_n-s_{n-1})}{n}$$

$$=s_n-\frac{s_1+s_2+\cdots+s_{n-1}}{n}$$

$$\to s-s=0. \qquad \text{Q.E.D.}$$

LEMMA 2. *If* $a_n\geqq 0$ *and* $a_n\to 0$ *in the sense of Cesaro, then also* $a_n^{1/2}\to 0$ *in the sense of Cesaro.*

Proof. We have

$$\frac{a_1^{1/2}+\cdots+a_n^{1/2}}{n}\leqq\frac{\left(n(a_1+\cdots+a_n)\right)^{1/2}}{n},$$

by Cauchy's inequality,

$$=\left[\frac{a_1+\cdots+a_n}{n}\right]^{1/2}\to 0. \qquad \text{Q.E.D.}$$

THEOREM 1. *If* $f(t)\in L^2[-1,1]$ *has Fourier cosine coefficients*

$$a_n=\int_{-1}^{1} f(t)\cos n\pi t\,dt,$$

then $n^{1/2}a_n\to 0$ *in the sense of Cesaro.*

Proof. By the Riesz–Fischer theorem, we have $\Sigma_1^\infty a_n^2<\infty$. Therefore, by Lemma 1, $na_n^2\to 0$ in the sense of Cesaro, and therefore, by Lemma 2, $n^{1/2}a_n\to 0$ in the sense of Cesaro. Q.E.D.

3. The result concerning $f(t)\in C[-1,1]$ depends on Fejer's theorem and a further lemma about Cesaro convergence.

LEMMA 3. *If* $\Sigma_1^\infty a_n$ *is convergent in the sense of Cesaro, then* $na_n\to 0$ *in the double Cesaro sense.*

Proof. Let

$$s_n=a_1+a_2+\cdots+a_n,$$

$$\sigma_n=\frac{s_1+s_2+\cdots+s_n}{n},$$

and suppose $\sigma_n\to\sigma$.

If

$$u_n=\frac{a_1+2a_2+\cdots+na_n}{n},$$

$$v_n=\frac{u_1+u_2+\cdots+u_n}{n},$$

then

$$u_n=\frac{s_1+2(s_2-s_1)+\cdots+n(s_n-s_{n-1})}{n}$$
$$=\frac{ns_n-s_1-s_2-\cdots-s_{n-1}}{n}$$
$$=s_n-\frac{n-1}{n}\sigma_{n-1}$$
$$=s_n-t_n,$$

where

$$t_n=\frac{n-1}{n}\sigma_{n-1}\to\sigma,$$

and hence

$$v_n=\sigma_n-\frac{t_1+t_2+\cdots+t_n}{n}$$
$$\to\sigma-\sigma=0. \qquad \text{Q.E.D.}$$

The counterexample $\Sigma_1^\infty(-1)^n$ shows that we cannot deduce $na_n\to 0$ in the single Cesaro sense under the same hypotheses.

THEOREM 2. *If $f(t)\in C[-1,1]$ has Fourier cosine coefficients*

$$a_n=\int_{-1}^{1}f(t)\cos n\pi t\,dt,$$

then $na_n\to 0$ in the double Cesaro sense.

Proof. By Fejer's theorem, $\Sigma_0^\infty a_n$ is convergent in the Cesaro sense (to $f(0)$). Hence the result follows from Lemma 3. Q.E.D.

Observe that we have only assumed $f(t)$ is continuous at $t=0$.

A uniform boundedness principle argument shows there exist $f(t)\in C[-1,1]$ for which na_n does not $\to 0$ in the single Cesaro sense. We need the following lemma.

LEMMA 4. *The variation of*

$$\frac{\sin\pi t+\sin 2\pi t+\cdots+\sin n\pi t}{n}=\frac{1}{2n}\left[\cot\tfrac{1}{2}\pi t-\frac{\cos(n+1/2)\pi t}{\sin(\pi t/2)}\right]$$

over $|t|\leqq 1$ tends to infinity as $n\to\infty$.

Proof. The graph of

$$\cot\tfrac{1}{2}\pi t-\frac{\cos(n+1/2)\pi t}{\sin(\pi t/2)}$$

lies between the graphs of $\cot\frac{1}{4}\pi t$ and $-\tan\frac{1}{4}\pi t$ over $0\leqq t\leqq 1$, touching each alternately at

$$t=\frac{1}{n+1/2},\frac{2}{n+1/2},\cdots,\frac{n}{n+1/2}.$$

Therefore

$$\operatorname*{var}_{0\leqq t\leqq 1}\left[\cot \tfrac{1}{2}\pi t-\frac{\cos(n+1/2)\pi t}{\sin(\pi t/2)}\right]$$

$$>2\left[\cot\left(\tfrac{1}{4}\frac{\pi}{n+1/2}\right)+\tan\left(\tfrac{1}{4}\frac{2\pi}{n+1/2}\right)+\cdots+\cot\left(\tfrac{1}{4}\frac{n\pi}{n+1/2}\right)\right]$$

if n odd,

$$>2\left(n+\tfrac{1}{2}\right)\left(1+\tfrac{1}{3}+\cdots+\tfrac{1}{n}\right)$$

using the inequality $\tan \tfrac{1}{4}\pi x<x$ for $0<x<1$,

$$\sim n\log n$$

as $n\to\infty$. Similarly for n even. Q.E.D.

COROLLARY. *If*

$$\phi_n(t)=\frac{d}{dt}\,\frac{\sin\pi t+\sin 2\pi t+\cdots+\sin n\pi t}{n},$$

then

$$\int_{-1}^{1}|\phi_n(t)|dt\to\infty$$

as $n\to\infty$.

Proof. Clearly,

$$\int_{-1}^{1}|\phi_n(t)|dt=\operatorname*{var}_{|t|\leqq 1}\frac{\sin\pi t+\sin 2\pi t+\cdots+\sin n\pi t}{n}.$$ Q.E.D.

Now suppose (for contradiction) that $na_n\to 0$ in the single Cesaro sense for every $f(t)\in C[-1,1]$. Then

$$\frac{a_1+2a_2+\cdots+na_n}{n}=\int_{-1}^{1}f(t)\,\frac{\cos\pi t+2\cos 2\pi t+\cdots+n\cos n\pi t}{n}\,dt$$

$$=\frac{1}{\pi}\int_{-1}^{1}f(t)\phi_n(t)\,dt$$

$$=T_nf,$$

where T_n is a linear functional on $C[-1,1]$ having

$$\|T_n\|=\frac{1}{\pi}\int_{-1}^{1}|\phi_n(t)|dt$$

which tends to infinity as $n\to\infty$, so $T_nf\to 0$ as $n\to\infty$ for every $f\in C[-1,1]$ contradicts the uniform boundedness principle.

4. We now show how ordinary convergence can be proved under the extra assumption that $a_n\geqq a_{n+1}\geqq 0$ for all n.

LEMMA 5. *If* $a_n \geqq a_{n+1} \geqq 0$ *and* $n^{1/2}a_n \to 0$ *in the sense of Cesaro, then* $n^{1/2}a_n \to 0$ *in the ordinary sense.*

Proof. Given any $\varepsilon > 0$, we have, for all large enough n,

$$\begin{aligned}\varepsilon &> \frac{a_1 + 2^{1/2}a_2 + \cdots + n^{1/2}a_n}{n} \\ &\geqq \frac{1 + 2^{1/2} + \cdots + n^{1/2}}{n} a_n \\ &> \frac{a_n}{n}\int_0^n t^{1/2}\,dt \\ &= \tfrac{2}{3}n^{1/2}a_n. \end{aligned}$$

Q.E.D.

LEMMA 6. *If* $a_n \geqq a_{n+1} \geqq 0$ *and* $na_n \to 0$ *in the double Cesaro sense, then* $na_n \to 0$ *in the ordinary sense.*

Proof. We have

$$u_n = \frac{a_1 + 2a_2 + \cdots + na_n}{n} \geqq \tfrac{1}{2}(n+1)a_n \geqq \tfrac{1}{2}na_n,$$

$$v_n = \frac{u_1 + u_2 + \cdots + u_n}{n} \geqq \frac{a_1 + 2a_2 + \cdots + na_n}{2n} = \tfrac{1}{2}u_n \geqq \tfrac{1}{4}na_n,$$

which gives the result. Q.E.D.

5. We now turn our attention to Fourier sine coefficients

$$b_n = \int_{-1}^{1} f(t)\sin n\pi t\,dt.$$

Theorem 1 goes through verbatim for sine coefficients. Theorem 2 also goes through, but a different proof is required since Fejer's theorem does not hold for sine coefficients. (Consider, e.g.,

$$f(t) = \sum_2^\infty \frac{\sin n\pi t}{n \log n}.)$$

LEMMA 7. *If* B *is a Banach space and* (T_n) *is a sequence of linear functionals on* B *such that* $\|T_n\|$ *is bounded over* n *and* $T_n f \to 0$ *as* $n \to \infty$ *for every* f *in some fundamental set in* B, *then* $T_n f \to 0$ *as* $n \to \infty$ *for every* f *in* B.

(We say a subset of B is *fundamental* if linear combinations of elements of the subset are dense in B.)

Proof. Let $f \in B$ and $\varepsilon > 0$ be given. Then we can choose a linear combination g of elements in the fundamental set such that $\|f - g\| < \varepsilon$. Now $T_n g \to 0$ as $n \to \infty$, so we can choose N such that $|T_n g| < \varepsilon$ for all $n > N$. Therefore

$$|T_n f| \leqq \|T_n\|\|f - g\| + |T_n g| < (M+1)\varepsilon$$

for all $n > N$, where $M = \sup_n \|T_n\|$. Q.E.D.

LEMMA 8. *The variation of*

$$\psi_n(t)=\frac{\sin^2(n\pi t/2)}{\sin^2(\pi t/2)}$$

over $|t|\leqq 1$ *is* $O(n^2)$ *as* $n\to\infty$.

Proof. The graph of $\psi_n(t)$ over $0\leqq t\leqq 1$ lies between the graph of $\operatorname{cosec}^2\frac{1}{2}\pi t$ and the horizontal axis, touching each alternately at

$$t=\frac{1}{n},\frac{2}{n},\cdots,1.$$

Hence, clearly,

$$\operatorname*{var}_{0\leqq t\leqq 1}\psi_n(t)<n^2+2\left(\operatorname{cosec}^2\left(\tfrac{1}{2}\frac{2\pi}{n}\right)+\operatorname{cosec}^2\left(\tfrac{1}{2}\frac{4\pi}{n}\right)+\cdots+\operatorname{cosec}^2\left(\tfrac{1}{2}\pi\right)\right)$$

if n is even,

$$<n^2+2n^2\left(\frac{1}{2^2}+\frac{1}{4^2}+\cdots+\frac{1}{n^2}\right),$$

using the inequality $\sin\frac{1}{2}\pi x>x$ for $0<x<1$,

$$<\left(1+\frac{\pi^2}{12}\right)n^2.$$

Similarly if n is odd. Q.E.D.

THEOREM 2′. *If* $f(t)\in C[-1,1]$ *has Fourier sine coefficients*

$$b_n=\int_{-1}^{1}f(t)\sin n\pi t\,dt,$$

then $nb_n\to 0$ *in the double Cesaro sense.*

Proof. If we write

$$u_n=\frac{b_1+2b_2+\cdots+nb_n}{n},$$

$$v_n=\frac{u_1+u_2+\cdots+u_n}{n},$$

and $T_nf=v_n$, then $T_nf\to 0$ as $n\to\infty$ for every f in the fundamental set

$$\tfrac{1}{2},\cos\pi t,\cos 2\pi t,\cdots,\sin\pi t,\sin 2\pi t,\cdots$$

in the Banach space $C[-1,1]$, and

$$\|T_n\|=\frac{1}{2\pi n^2}\operatorname*{var}_{|t|\leqq 1}\psi_{n+1}(t)$$

is bounded over n. Q.E.D.

If instead we take $T_nf=u_n$, we have

$$\|T_n\|=\frac{1}{2\pi n}\operatorname*{var}_{|t|\leqq 1}\frac{\sin(n+1/2)\pi t}{\sin(\pi t/2)}$$

which tends to infinity as $n \to \infty$ since

$$n \log n = O\left[\operatorname*{var}_{|t| \leqq 1} \frac{\sin(n+1/2)\pi t}{\sin(\pi t/2)} \right]$$

by a similar argument to that used in the proof of Lemma 4. So it follows by the uniform boundedness principle that there exist $f \in C[-1,1]$ for which nb_n does not $\to 0$ in the single Cesaro sense.

6. If $f(t) \in AC[-1,1]$ is absolutely continuous over the interval $[-1,1]$, then $f(t)$ has a derivative $f'(t) \in L^1[-1,1]$ and the fundamental theorem of calculus holds. If also $f(1)=f(-1)$, then we have

$$\begin{aligned} a_n' &= \int_{-1}^{1} f'(t) \cos n\pi t \, dt \\ &= f(t) \cos n\pi t \Big|_{-1}^{1} + n\pi \int_{-1}^{1} f(t) \sin n\pi t \, dt \\ &= n\pi b_n, \\ b_n' &= \int_{-1}^{1} f'(t) \sin n\pi t \, dt \\ &= f(t) \sin n\pi t \Big|_{-1}^{1} - n\pi \int_{-1}^{1} f(t) \cos n\pi t \, dt \\ &= -n\pi a_n. \end{aligned}$$

Hence, always assuming $f(1)=f(-1)$, we have $na_n \to 0$ and $nb_n \to 0$ in general (see [1, p. 24]), $n^{3/2}a_n \to 0$ and $n^{3/2}b_n \to 0$ in the sense of Cesaro if $f'(t) \in L^2[-1,1]$, and $n^2 a_n \to 0$, $n^2 b_n \to 0$ in the double Cesaro sense if $f(t) \in C^1[-1,1]$. It follows, as in §4, that the second two pairs of results hold for convergence in the ordinary sense if the Fourier coefficients are real, positive and decreasing. Observe that the condition $f(1)=f(-1)$ is not needed for the results concerning cosine coefficients since the proof of $b_n' = -n\pi a_n$ does not require this condition.

The second two results for monotonic cosine coefficients can be obtained as corollaries of theorems about the order of magnitude of the eigenvalues of certain integral equations. In fact, if real, even $f(t) \in L^1[-1,1]$ is extended to $[-2,2]$ by requiring it to be periodic with period 2, then

$$T\phi(x) = \int_{-1}^{1} f(x-t)\phi(t) \, dt$$

defines a compact symmetric operator T on the Hilbert space $L^1[-1,1]$ whose eigenvalues are $(a_n)_{n \geqq 0}$ (multiplicity one for $n=0$, two for $n \geqq 1$). H. Weyl has shown (see [4]) that for an operator

$$T\phi(x) = \int_{a}^{b} K(x,t)\phi(t) \, dt$$

with a general symmetric C^1 kernel $K(x,t)$ defined on the square $[a,b]^2$, the eigenvalues (λ_n) satisfy $n^{3/2}\lambda_n \to 0$ when arranged in descending order of modulus. It can be shown that this result still holds when $K(x,t)$ is assumed to be absolutely continuous in

each variable for each fixed value of the other variable and the partial derivatives are in $L^2[a,b]^2$ (see [3]). It can also be shown that, if the eigenvalues are positive and $K(x,t) \in C^1[a,b]^2$, then $n^2\lambda_n \to 0$ (see [2]).

Similar results hold with correspondingly higher powers of n for functions $f(t)$ with higher order derivatives.

REFERENCES

[1] Y. KATNZELSON, *An Introduction to Harmonic Analysis*. Jerusalem, 1968, John Wiley, New York.

[2] J. B. READE, *Eigenvalues of smooth kernels*, Math. Proc. Camb. Phil. Soc., 95 (1984), pp. 135–140.

[3] _____, *Eigenvalues of positive definite kernels*, this Journal, 14 (1983), pp. 152–157.

[4] H. WEYL, *Das asymptotische Verteilungsgestz der Eigenwerte linearer partieller Differentialgleichungen*, Math. Ann., 71 (1912), pp. 441–479.

SIAM J. MATH. ANAL.
Vol. 17, No. 2, March 1986

015

GENERALIZATION OF ZEMANIAN SPACES OF GENERALIZED FUNCTIONS WHICH HAVE ORTHONORMAL SERIES EXPANSIONS*

STEVAN PILIPOVIĆ†

Abstract. We define the space of generalized functions $\exp_p \mathscr{A}'$, $p \in N$. If $f \in \exp_p \mathscr{A}'$, then f can be uniquely expanded into a series of the form

$$(*) \qquad f = \sum_{n=0}^{\infty} a_n \psi_n,$$

where $\{\psi_n\}_{n=0}^{\infty}$ is an orthonormal base of the corresponding space $L^2(I)$, I is an interval in R, $\{b_n\}_{n=0}^{\infty}$ is a sequence of complex numbers, such that for some $k \in N$,

$$(**) \qquad b_n = O\left(\left(\underbrace{\exp \exp \cdots \exp}_{p} \tilde{\lambda}_n\right)^k\right), \qquad n = 0, 1, \cdots;$$

$\{\lambda_n\}_{n=0}^{\infty}$ is a sequence of eigenvalues of a corresponding operator $\mathscr{R}$ ($\mathscr{R}\psi_n = \lambda_n \psi_n$), and $\tilde{\lambda}_n = |\lambda_n|$ if $\lambda_n \neq 0$ and $\tilde{\lambda}_n = 1$ if $\lambda_n = 0$.

The series in (*) converges in the sense of weak topology in $\exp_p \mathscr{A}'$. Conversely, if a sequence $\{b_n\}_{n=0}^{\infty}$ satisfies (**) for some $k \in N$, a unique element in $\exp_p \mathscr{A}'$ is defined by the series in (*).

We give representation theorems for elements from $\exp_p \mathscr{A}'$, $p = 1, 2, \cdots$, which show that spaces $\exp \mathscr{A}$; $\exp_2 \mathscr{A}', \cdots$ are natural generalizations of the space $\mathscr{A}'$ from [8].

1. Introduction and notation. In his book [8], Zemanian presented remarkable methods for constructing spaces of generalized functions which correspond to integral transformations and differential operators.

In this paper we shall generalize results from [8, Chap. 9], so we shall use notations from this monograph.

Let I be an open interval of the real line R, and let $L^2(I)$ be the space of square integrable functions with the usual norm. We denote by $\mathscr{R}$ a linear differential self-adjoint operator of the form

$$\mathscr{R} = \theta_0 D^{n_1} \theta_1 D^{n_2} \cdots D^{n_\nu} \theta_\nu$$

such that

$$\mathscr{R} = \bar{\theta}_\nu (-D)^{n_\nu} \cdots (-D)^{n_2} \bar{\theta}_1 (-D)^{n_1} \bar{\theta}_0,$$

where $D = d/dx$, $\{\eta_k\}_{k=1}^{\nu}$ are nonnegative integers, $\{\theta_k\}_{k=0}^{\nu}$ are smooth functions on I without zeros, and $\bar{\theta}_k$ are complex conjugates of θ_k, $k = 0, 1, \cdots, \nu$. We suppose that there exist a sequence of real numbers $\{\lambda_n\}_{n=0}^{\infty}$ and a sequence of smooth functions $\{\psi_n\}_{n=0}^{\infty}$ such that $\mathscr{R}\psi_n = \lambda_n \psi_n$. Furthermore we suppose that the sequence $\{|\lambda_n|\}_{n=0}^{\infty}$ monotonically tends to infinity and that $\{\psi_n\}$ forms an orthonormal base of $L^2(I)$.

* Received by the editors January 23, 1984, and in revised form June 15, 1984.
† University of Novi Sad, Institute of Mathematics, Novi Sad, Yugoslavia.

If $\{\alpha_n\}_{n=0}^{\infty}$ is a sequence of complex numbers different from zero, we denote by $S(\alpha_n)$ and $S^x(\alpha_n)$ the sequence spaces defined in the following way (see [2]):

$$\{a_n\}_{n=0}^{\infty} \in S(\alpha_n) \text{ iff for every } k \in N_0,\ \sum_{n=0}^{\infty} |a_n|^2 |\alpha_n|^{2k} < \infty,$$

$$\{b_n\}_{n=0}^{\infty} \in S^x(\alpha_n) \text{ iff for some } k \in N_0,\ \sum_{n=0}^{\infty} |b_n|^2 |\alpha_n|^{-2k} < \infty$$

($N_0 = N \cup \{0\}$).

Zemanian proved in [8, Chap. 9] that there exists a bijection between the space of generalized functions $\mathscr{A}'$ whose elements may be uniquely expanded into a series and the space $S^x(\tilde{\lambda}_n)$ where $\tilde{\lambda}_n = |\lambda_n|$ if $\lambda_n \neq 0$ and $\tilde{\lambda}_n = 1$ if $\lambda_n = 0$, $n = 0, 1, \cdots$.

In this paper we shall investigate the space of generalized functions $\exp\mathscr{A}'$ which contains $\mathscr{A}'$ as a proper subspace. The elements of the space $\exp\mathscr{A}'$ have unique orthonormal expansions of the form

$$\exp\mathscr{A}' \ni f = \sum_{n=0}^{\infty} b_n \psi_n, \qquad b_n = \langle f, \bar{\psi}_n \rangle, \qquad n \in N_0,$$

where this series converges weakly in $\exp\mathscr{A}'$. We shall show that there exists a bijection between the spaces $\exp\mathscr{A}'$ and $S^x(\exp\tilde{\lambda}_n)$:

$$\exp\mathscr{A}' \ni f = \sum_{n=0}^{\infty} b_n \psi_n \leftrightarrow \{b_n\} \in S^x(\exp\tilde{\lambda}_n).$$

We shall give the representation theorem for the element of the space $\exp\mathscr{A}'$ Moreover, we shall investigate further generalizations of the spaces $\exp\mathscr{A}'$, spaces $\exp\exp\mathscr{A}'$, $\exp\exp\exp\mathscr{A}', \cdots$, for which we shall give representation theorems.

We shall briefly show in §5, Remark 2 that, similar to [8, 9.7], our theory can be applied in solving a class of differential equations of infinite order.

If we use the approach from the paper [6], we can construct the most general space of generalized functions whose elements have orthonormal expansion of the form $f = \sum_{n=0}^{\infty} b_n \psi_n$ without any condition on the coefficients b_n, $n \in N_0$.

In this paper our intention is to more precisely characterize generalized functions which have unique orthonormal expansions; for that purpose the spaces $\exp_p\mathscr{A}'$ are more convenient.

We notice that the generalizations of the spaces $\mathscr{A}'$ given in this paper can be done for the spaces $\tilde{\mathscr{A}}'$ from [7] without difficulty.

2. Spaces $\exp\mathscr{A}$ and $\exp\mathscr{A}'$. We denote by $\exp\mathscr{A}$ the subspace of $L^2(I)$ defined in the following way:

$$\phi \overset{2}{=} \sum_{n=0}^{\infty} a_n \psi_n \in \exp\mathscr{A} \quad \text{iff } \{a_n\} \in S(\exp\tilde{\lambda}_n)$$

($\overset{2}{=}$ means equality in the L^2 sense. If there is no possibility for misinterpretation, we shall put $=$ instead of $\overset{2}{=}$).

In the space $\exp\mathscr{A}$ we introduce the sequence of norms

$$\gamma_k(\phi) := \sqrt{\sum_{n=0}^{\infty} |a_n|^2 (\exp\tilde{\lambda}_n)^{2k}}, \qquad k \in N_0.$$

This sequence of norms defines the usual topology in $S(\exp\tilde{\lambda}_n)$ [2]. Thus the spaces $\exp\mathscr{A}$ and $S(\exp\tilde{\lambda}_n)$ are isomorphic.

If we denote by $(\exp\mathscr{A})_k$ the subspace of $L^2(I)$ such that

$$\phi \in (\exp\mathscr{A})_k \quad \text{iff} \quad \gamma_k(\phi) < \infty,$$

it is easy to show that $(\exp\mathscr{A})_k$, $k \in N_0$, are Banach spaces. The sequence of norms $\{\gamma_k\}_{k=0}^{\infty}$ is pairwise, compatible [1, p. 13] and monotone ($\gamma_p(\phi) \leqq \gamma_{p+1}(\phi)$, $p \in N_0$). Since $\exp\mathscr{A}$ is a dense subspace of $(\exp\mathscr{A})_k$ according to the norm γ_k, from [1, p. 36] we have

$$\exp\mathscr{A}' = \bigcup_{k=0}^{\infty} (\exp\mathscr{A})_k'. \tag{1}$$

PROPOSITION 1. (i) *The space* $\exp\mathscr{A}$ *is a dense subspace of* $\mathscr{A}$ and $\mathscr{E}(I)$. *Moreover, the inclusion mappings of* $\exp\mathscr{A}$ *into* $\mathscr{A}$ *or* $\mathscr{E}(I)$ *are continuous.* ($\mathscr{E}(I)$ *is the space* $C^{\infty}(I)$ *with the usual topology.*)

(ii) *The space* $\exp\mathscr{A}$ *is nuclear if for some* $k \in N_0$, $\sum_{n=0}^{\infty} (\exp\tilde{\lambda}_n)^{-2k} < \infty$.

(iii) *If* $\phi \in \exp\mathscr{A}$, $\phi = \sum_{n=0}^{\infty} a_n \psi_n$ *then for any* $k \in N_0$, $\mathscr{R}^k\phi = \sum_{n=0}^{\infty} \lambda_n^k a_n \psi_n \in \exp\mathscr{A}$, *where the series converges in the sense of convergence in* $\exp\mathscr{A}$.

Proof. (i) If $\phi = \sum_{n=0}^{\infty} a_n\psi_n \in \mathscr{A}$, then $\{\sum_{n=0}^{\nu} a_n\psi_n\}$ is a sequence from $\exp\mathscr{A}$ which in the sense of convergence in $\mathscr{A}$ converges to ϕ. If

$$\phi_p = \sum_{n=0}^{\infty} a_{n,p}\psi_n, \qquad p \in N,$$

is a sequence from $\mathscr{A}$ and $\phi = \sum_{n=0}^{\infty} a_n\psi_n \in \mathscr{A}$, then $\phi_p \to \phi$ in $\mathscr{A}$ iff for every $k \in N_0$, $\sum_{n=0}^{\infty} |a_{n,p} - a_n|^2 \tilde{\lambda}_n^{2k} \to 0$. So it is clear that the inclusion mapping (i): $\exp\mathscr{A} \to \mathscr{A}$ is continuous. Since $\mathscr{A}$ is a dense subspace of $\mathscr{E}(I)$ and the inclusion mapping i: $\mathscr{A} \to \mathscr{E}(I)$ is continuous ([8, Lem. 9.3.4]), it follows that $\exp\mathscr{A}$ is a dense subspace of $\mathscr{E}(I)$ and the inclusion mapping i: $\exp\mathscr{A} \to \mathscr{E}(I)$ is continuous.

(ii) This follows from [3] because the space $S(\exp\tilde{\lambda}_n)$ is equal to the space $\lambda(\exp\tilde{\lambda}_n) = \{\{x_n\}_{n=0}^{\infty};\ \sum_{n=0}^{\infty} |x_n| (\exp\tilde{\lambda}_n)^k < \infty$ for every $k \in N_0\}$ iff for some $k \in N_0$, $\sum_{n=0}^{\infty} (\exp\tilde{\lambda}_n)^{-k} < \infty$ holds ([5]).

(iii) Since this assertion holds for elements from $\mathscr{A}$, it is easy to prove that this holds for elements from $\exp\mathscr{A}$.

THEOREM 2. (i) *If* $\phi \in \exp\mathscr{A}$, *then for every* $k \in N_0$,

$$\theta_k(\phi) := \sum_{n=0}^{\infty} \frac{k^n}{n!} \|\mathscr{R}^n\phi\|_2 < \infty.$$

(ii) *The sequences of norms* $\{\theta_k\}$ *and* $\{\gamma_k\}$ *in* $\exp\mathscr{A}$ *are equivalent.*

Proof. (i) Let $\phi = \sum_{n=0}^{\infty} a_n\psi_n \in \exp\mathscr{A}$. Since

$$\|\mathscr{R}^k\phi\|_2 = \sqrt{\sum_{p=0}^{\infty} |a_p|^2 \lambda_p^{2k}},$$

we have (with $m>k$ and some $C>0$)

$$\theta_k(\phi)=\sum_{n=0}^{\infty}\frac{k^n}{n!}\sqrt{\sum_{p=0}^{\infty}|a_p|^2\lambda_p^{2n}}=\sum_{n=0}^{\infty}\sqrt{\sum_{p=0}^{\infty}\frac{k^{2n}}{n!^2}|a_p|^2\lambda_p^{2n}}$$

$$\leqq C\sqrt{\sum_{n=0}^{\infty}\sum_{p=0}^{\infty}\left(\frac{m^n}{n!}|a_p|\tilde{\lambda}_p^n\right)^2}=C\sqrt{\sum_{p=0}^{\infty}|a_p|^2\sum_{n=0}^{\infty}\left(\frac{m^n}{n!}\tilde{\lambda}_p^n\right)^2}$$

$$\leqq C\sqrt{\sum_{p=0}^{\infty}|a_p|^2\left(\sum_{n=0}^{\infty}\frac{m^n}{n!}\tilde{\lambda}_p^n\right)}=C\sqrt{\sum_{p=0}^{\infty}|a_p|^2(\exp\tilde{\lambda}_p m)}$$

$$\leqq C\gamma_m(\phi)<\infty.$$

(ii) In (i) we proved that $\theta_k(\phi)\leqq C\gamma_m(\phi)$, so we have to prove that for a given $k\in N_0$ there exist $C>0$ and $m\in N_0$ such that

$$\gamma_k(\phi)\leqq C\theta_m(\phi).$$

Since finitely many λ_n may be equal to zero, we have (with $m>r>k$ and suitable $C_{11}C_{21}C$)

$$\gamma_k(\phi)=\sqrt{\sum_{n=0}^{\infty}|a_n|^2(\exp\tilde{\lambda}_n k)^2}\leqq C_1\sqrt{\sum_{n=0}^{\infty}|a_n|^2\left(\sum_{p=0}^{\infty}\frac{(|\lambda_n|k)^p}{p!}\right)^2}$$

$$\leqq C_2\sqrt{\sum_{n=0}^{\infty}|a_n|^2\sum_{p=0}^{\infty}\frac{|\lambda_n|^{2p}r^{2p}}{p!^2}}=C_2\sqrt{\sum_{p=0}^{\infty}\frac{r^{2p}}{p!^2}\sum_{n=0}^{\infty}|a_n|^2|\lambda_n|^{2p}}$$

$$\leqq C\sup_p\left\{\frac{m^p}{p!}\sqrt{\sum_{n=0}^{\infty}|a_n|^2|\lambda_n|^{2p}}\right\}\leqq C\sum_{p=0}^{\infty}\frac{m^p}{p!}\sqrt{\sum_{n=0}^{\infty}|a_n|^2|\lambda_n|^{2p}}=C\theta_m(\phi).$$

So the proof is complete.

If we denote by $\widetilde{(\exp\mathscr{A})}_k$, $k\in N_0$, a subspace of $\mathscr{A}$ such that $\phi\in(\exp\mathscr{A})_k$ iff $\theta_k(\phi)<\infty$, by the standard arguments one can prove:

PROPOSITION 3. *The space* $\widetilde{(\exp\mathscr{A})}_k$, $k\in N_0$, *is a Banach space.*

The sequence of norms $\{\theta_k\}$ monotonically increases and any two norms from this sequence are consistent.

If $\phi\overset{2}{=}\sum_{n=0}^{\infty}a_n\psi_n$ is an arbitrary element from $(\exp\mathscr{A})_k$, then the sequence $\{\sum_{n=0}^{\nu}a_n\psi_n\}_{\nu=0}^{\infty}$ from $\exp\mathscr{A}$ converges to ϕ in the sense of convergence in $\widetilde{(\exp\mathscr{A})}_k$.

So $\widetilde{(\exp\mathscr{A})}_k$ is a completion of $\exp\mathscr{A}$ according to the norm θ_k, and we have:

PROPOSITION 4. $\exp\mathscr{A}'=\bigcup_{k=0}^{\infty}\widetilde{(\exp\mathscr{A})}_k'$.

3. Representation theorems. [8, Thms. 9.5.1 and 9.6.1] directly imply:

THEOREM 5. (i) *If* $f\in\exp\mathscr{A}'$, *then there exists a sequence of complex numbers* $\{b_n\}_{n=0}^{\infty}$ *such that*

$$f=\sum_{n=0}^{\infty}b_n\psi_n,\qquad b_n=\langle f,\bar{\psi}_n\rangle,\qquad n\in N_0,\tag{2}$$

where the series converges weakly in $\exp\mathscr{A}'$.

(ii) *The series on the right side of* (2) *converges in* $\exp\mathscr{A}'$ *iff there exists* $r\in N_0$ *such that* $\sum_{n=0}^{\infty}|b_n|^2\exp(-2r\tilde{\lambda}_n)<\infty$ (*that is*, $\{b_n\}\in S^x(\exp\tilde{\lambda}_n)$).

We shall prove:

THEOREM 6. *If* $f\in\exp\mathscr{A}'$, *there exists a sequence* $\{f_n\}_{n=0}^{\infty}$ *from* $L^2(I)$ *and* $k\in N_0$ *such that*

$$f=\sum_{n=0}^{\infty}\frac{k^n}{n!}\mathscr{R}^n f_n \tag{3}$$

and

$$\sup_{n\in N_o}\|f_n\|_2<\infty. \tag{4}$$

Conversely, if a sequence $\{f_n\}$ *from* $L^2(I)$ *satisfies* (4), *with the series on the right side of* (3) *a unique element from* $\exp\mathscr{A}'$ *is defined.*

Proof. For the proof we shall use an idea from [9] (see also [4]). If $f\in\exp\mathscr{A}'$, then from Proposition 4 it follows that f may be extended from $\exp\mathscr{A}$ onto $(\widetilde{\exp\mathscr{A}})_k$ for some $k\in N_0$, to become an element of $(\widetilde{\exp\mathscr{A}})'_k$. We denote this element from $(\widetilde{\exp\mathscr{A}})'_k$ again by f. So we shall give a representation of any element from $(\widetilde{\exp\mathscr{A}})'_k$, $k\in N_0$.

We denote by Γ the subspace of the Tikhonov product $\Pi_{p=0}^{\infty}L^2(I)$ defined in the following way:

$$\{f_n\}_{n=0}^{\infty}\in\Gamma \quad\text{iff}\quad \|\{f_n\}\|_\Gamma:=\sum_{n=0}^{\infty}\|f_n\|_2<\infty.$$

The mapping α from $(\widetilde{\exp\mathscr{A}})_k$ to Γ is defined by

$$\alpha(\phi)=\left\{\frac{k^n}{n!}\mathscr{R}^n\phi\right\}_{n=0}^{\infty}$$

is an isometry of $(\widetilde{\exp\mathscr{A}})_k$ and of $((\widetilde{\exp\mathscr{A}})_k)\subset\Gamma$. We define a continuous linear functional $\tilde{f}$ on $\alpha((\widetilde{\exp\mathscr{A}})_k)$ by

$$\langle\tilde{f},\psi\rangle=\langle f,\alpha^{-1}(\psi)\rangle\psi\in\alpha\left(\left(\widetilde{\exp\mathscr{A}}\right)_k\right).$$

From the Hahn–Banach theorem it follows that $\tilde{f}$ may be extended onto Γ continuously and linearly. We denote this extension by F. It is known that if $F\in\Gamma'$, then there exists a sequence of functions $\{f_n\}$ from $L^2(I)$ such that

$$\langle F,\{\psi_n\}\rangle=\sum_{n=0}^{\infty}\int_I f_n(x)\psi_n(x)\,dx,\ \{\psi_n\}\in\Gamma,\quad\text{and}\quad\sup_{n\in N_0}\|f_n\|_2<\infty.$$

This means that if $\phi\in\exp\mathscr{A}$, then in the sense of weak convergence in $\exp\mathscr{A}'$, we have

$$\begin{aligned}\langle f,\phi\rangle&=\sum_{n=0}^{\infty}\int_I f_n(x)\frac{k^n}{n!}\mathscr{R}^k\phi(x)\,dx=\sum_{n=0}^{\infty}\frac{k^n}{n!}\langle f_n,\mathscr{R}^k\phi\rangle\\&=\sum_{n=0}^{\infty}\left\langle\frac{k^n}{n!}\mathscr{R}^k f_n,\phi\right\rangle=\left\langle\sum_{n=0}^{\infty}\frac{k^n\mathscr{R}^k f_n}{n!},\phi\right\rangle.\end{aligned} \tag{5}$$

It follows that f is of the form (3) such that (4) holds.

4. Spaces $\exp_p \mathscr{A}$ and $\exp_p \mathscr{A}'$ ($p \geq 2$). If $\phi \in \exp\mathscr{A}$ then the series $\sum_{n=0}^{\infty}(k^n/n!)\mathscr{R}^k\phi$ converges in $L^2(I)$. This follows from Theorem 2 (i). Let us denote by

$$E_1^k = (\exp\mathscr{R})^k = \exp k\mathscr{R}, \qquad k \in N_0,$$

the linear differential operator from $\exp\mathscr{A}$ to $L^2(I)$ defined by

$$\phi \mapsto (\exp k\mathscr{R})\phi := \sum_{n=0}^{\infty} \frac{k^n}{n!}\mathscr{R}^n\phi, \qquad \phi \in \exp\mathscr{A}.$$

We have

PROPOSITION 7. *The operator E_1^k, $k \in N_0$, is the continuous operator of the space $\exp\mathscr{A}$ into the same space.*

Proof. If $\phi = \sum_{n=0}^{\infty} a_n\psi_n \in \exp\mathscr{A}$ and $E_1^k\phi = \sum_{n=0}^{\infty} c_n\psi_n \in L^2(I)$, we have

$$c_p = \left\langle \sum_{n=0}^{\infty} \frac{k^n}{n!}\mathscr{R}^n\phi, \bar{\psi}_p \right\rangle = \sum_{n=0}^{\infty} \frac{k^n}{n!}\lambda_p^n\langle\phi, \bar{\psi}_p\rangle = a_p \exp(k\lambda_p), \qquad p \in N_0. \tag{6}$$

From (6) it easily follows that $E_1^k\phi \in \exp\mathscr{A}$ and that the mapping E_1^k: $\exp\mathscr{A} \mapsto \exp\mathscr{A}$ is continuous.

We denote by $\exp\exp\mathscr{A}$ the subspace of $\exp\mathscr{A}$ defined in the following way

$$\phi = \sum_{n=0}^{\infty} a_n\psi_n \in \exp\exp\mathscr{A}$$

iff for every $k \in N_0$,

$${}_2\gamma_k(\phi) := \sqrt{\sum_{n=0}^{\infty} |a_n|^2 (\exp\exp\tilde{\lambda}_n)^{2k}} < \infty \quad (\text{that is, } \{a_n\} \in S(\exp\exp\tilde{\lambda}_n)).$$

If $\phi = \sum_{n=0}^{\infty} a_n\psi_n \in \exp\mathscr{A}$, then $\{\sum_{n=0}^{\nu} a_n\psi_n\}_{\nu=0}^{\infty}$ is a sequence from $\exp\exp\mathscr{A}$ which converges to ϕ in $\exp\mathscr{A}$. This implies that $\exp\exp\mathscr{A}$ may be continuously embedded into $\exp\mathscr{A}$ and $\mathscr{E}(I)$ as a dense subspace.

The sequence of norms $\{{}_2\gamma_k\}_{k=0}^{\infty}$ is equivalent with the following sequence of norms on $\exp\exp\mathscr{A}$:

$${}_2\theta_k(\phi) := \sum_{n=0}^{\infty} \frac{k^n}{n!} \sum_{m=0}^{\infty} \frac{n^m}{m!} \|\mathscr{R}^m\phi\|_2.$$

This may be proved similarly as in Theorem 2.

We may construct in the same way the spaces $\exp\exp\exp\exp\mathscr{A}$, $\exp\exp\exp\mathscr{A}, \cdots$, defined respectively by the norms $\{{}_3\gamma_k\}$, $\{{}_4\gamma_k\}, \cdots$, where

$${}_p\gamma_k(\phi) := \sqrt{\sum_{n=0}^{\infty} |a_n|^2 \Big(\underbrace{\exp\exp\cdots\exp}_{p}\tilde{\lambda}_n\Big)^{2k}}, \qquad k \in N_0, \qquad p = 3, 4, \cdots,$$

is a norm on

$$\Big(\underbrace{\exp\cdots\exp}_{p}\mathscr{A}\Big).$$

Let us denote the latter space $\exp_p\mathscr{A}$. The operator

$$E_p^k=\Big(\underbrace{\exp\exp\cdots\exp}_{p}\mathscr{R}\Big)^k,\qquad k\in N_0,\ p\geqq 2,$$

on $\exp_p\mathscr{A}$, $p\geqq 2$, defined by

$$E_p^k\phi=\sum_{m_1=0}^{\infty}\frac{k^{m_1}}{m_1!}\sum_{m_2=0}^{\infty}\frac{m_1^{m_2}}{m_2!}\cdots\sum_{m_p=0}^{\infty}\frac{m_{p-1}^{m_p}}{m_p!}\mathscr{R}^{m_p}\phi,\qquad \phi\in\exp_p\mathscr{A},$$

continuously maps $\exp_p\mathscr{A}$ into $\exp_p\mathscr{A}$.

Similarly as for the space $\exp\mathscr{A}$ one can prove:

THEOREM 8. (i) *The space* $\exp_p\mathscr{A}$, $p\geqq 2$, *may be continuously embedded into the spaces* $\exp_{p-1}\mathscr{A},\cdots,\exp\mathscr{A},\mathscr{E}(I)$.

(ii) *The sequence of norms* $\{{}_p\gamma_k\}_{k=0}^{\infty}$ *on* $\exp\mathscr{A}$, $p\geqq 2$, *is equivalent to the sequence of norms*

$${}_p\theta_k(\phi):=\sum_{m_1=0}^{\infty}\frac{k^{m_1}}{m_1!}\sum_{m_2=0}^{\infty}\frac{m_1^{m_2}}{m_2!}\cdots\sum_{m_p=0}^{\infty}\frac{m_{p-1}^{m_p}}{m_p!}\|\mathscr{R}^{m_p}\phi\|_2,\qquad k\in N_0,\qquad p\geqq 2.$$

This means that $\mathscr{E}'(I),\mathscr{A}',\cdots,\exp_{p-1}\mathscr{A}'$ *are subspaces of* $\exp_p\mathscr{A}'$.

Now, we shall give the generalizations of Theorems 5 and 6.

THEOREM 9. (i) *If* $f\in\exp_p\mathscr{A}'$, *then there exists a sequence of complex numbers* $\{b_n\}$ *such that*

$$f=\sum_{n=0}^{\infty}b_n\psi_n,\qquad b_n=\langle f,\bar{\psi}_n\rangle,\qquad n\in N_0,\tag{7}$$

where the series converges weakly in $\exp_p\mathscr{A}$.

(ii) *The series on the right side of* (7) *converges in* $\exp_p\mathscr{A}'$ *iff there exists an* $r\in N_0$ *such that*

$$\sum_{n=0}^{\infty}|b_n|^2\exp\Big(-2r\Big(\underbrace{\exp\cdots\exp}_{p-1}\lambda_n\Big)\Big)<\infty\ \Big(\{b_n\}\in S^x\Big(\underbrace{\exp\cdots\exp}_{p}\tilde{\lambda}_n\Big)\Big).$$

THEOREM 10. *If* $f\in\exp_p\mathscr{A}$, *there exist a sequence* $\{f_{(m_1,\cdots,m_p)}\}$, $(m_1,\cdots,m_p)\in N_0^p$ *from* $L^2(I)$ *and a* $k\in N_0$ *such that*

$$f=\sum_{m_1=0}^{\infty}\frac{k^{m_1}}{m_1!}\sum_{m_2=0}^{\infty}\frac{m_1^{m_2}}{m_2!}\cdots\sum_{m_p=0}^{\infty}\frac{m_{p-1}^{m_p}}{m_p!}\mathscr{R}^{m_p}f_{(m_1,\cdots,m_p)}\tag{8}$$

and

$$\sup\Big\{\big\|f_{(m_1,\cdots,m_p)}\big\|_2;\ (m_1,m_2,\cdots,m_p)\in N_0^p\Big\}<\infty.\tag{9}$$

Conversely, if a sequence $\{f_{(m_1,\cdots,m_p)}\}$ *from* $L^2(I)$ *satisfies* (9), *a unique element from* $\exp_p\mathscr{A}'$ *is defined by the series on the right of* (8).

5. Remarks.

Remark 1. It is clear that if $\{\mu_n\}_{n=0}^{\infty}$ is a sequence of complex numbers such that

$$\mu_n = O\left(\left(\underbrace{\exp\cdots\exp}_{p}\tilde{\lambda}_n\right)^k\right) \quad \text{for some } k\in N,$$

then by

$$\sum_{n=0}^{\infty} a_n\psi_n \mapsto \sum_{n=0}^{\infty} a_n\mu_n\psi_n, \tag{10}$$

a linear continuous operator of multiplier type from $\mathscr{A}'$ into $\exp_p\mathscr{A}'$ or from $\exp_m\mathscr{A}'$ into $\exp_{m+p}\mathscr{A}'$ is defined. (We suppose in (10) that the series converges in the weak sense of the corresponding spaces of generalized functions.)

Remark 2. In the preceding section we defined differential operators of infinite orders E_p^k, $p\in N$, $k\in N$. As in [8, 9.7], one can easily show that the differential equation of the form

$$\left(P_0(\mathscr{R})+P_1(E_1)+\cdots+P_p(E_p)\right)f=g,$$

where $g\in\exp_p\mathscr{A}'$ and $P_0, P_1,\cdots,P_p$ are arbitrary polynomials, has a solution in the space $\exp_p\mathscr{A}'$.

REFERENCES

[1] I. M. GEL'FAND AND G. E. SHILOV, *Generalized Functions* 2, *Spaces of Fundamental and Generalized Functions*, Academic Press, New York, 1968.

[2] G. KÖTHE, *Topological Vector Spaces* I, Springer, Berlin, 1969.

[3] _____, *Nuclear sequence spaces*, Math. Balkan., 1 (1971), pp. 144–146.

[4] H. KOMATSU, *Ultradistributions*, I, *Structure theorems and a characterization*, J. Fac. Sci., Univ. Tokyo Sec. IA, 20 (1973), pp. 25–105.

[5] S. PILIPOVIĆ, *The kernel theorem for some spaces*, Rev. Res. Fac. Sci.-Univ. Novi Sad., 10 (1980), pp. 55–61.

[6] A. SZAZ, *Periodic generalized functions*, Publ. Math. (Debrecen) 25(1978), pp. 227–235.

[7] G. WALTER AND P. NEVAI, *Series of orthogonal polynomials as boundary values*, this Journal, 12 (1981), pp. 502–513.

[8] A. H. ZEMANIAN, *Generalized Integral Transformations*, Interscience, New York, 1968.

[9] T. YAMANAKA, *Note on some functions in Gel'fand and Shilov's theory of generalized functions*, Comment. Math. Univ. St. Paul., 9 (1961), pp. 1–6.

SIAM J. MATH. ANAL.
Vol. 17, No. 2, March 1986

016

ANALYTICITY SPACES OF SELF-ADJOINT OPERATORS SUBJECTED TO PERTURBATIONS WITH APPLICATIONS TO HANKEL INVARIANT DISTRIBUTION SPACES*

S. J. L. VAN EIJNDHOVEN† AND J. DE GRAAF†

Abstract. A new theory of generalized functions has been developed by one of the authors (de Graaf). In this theory the analyticity domain of each positive self-adjoint unbounded operator $\mathscr{A}$ in a Hilbert space X is regarded as a test space denoted by $\mathscr{S}_{X,\mathscr{A}}$. In the first part of this paper, we consider perturbations $\mathscr{P}$ on $\mathscr{A}$ for which there exists a Hilbert space Y such that $\mathscr{A}+\mathscr{P}$ is a positive self-adjoint operator in Y. In particular, we investigate for which perturbations $\mathscr{P}$ and for which $\nu>0$, $\mathscr{S}_{X,\mathscr{A}^\nu}\subset\mathscr{S}_{Y,(\mathscr{A}+\mathscr{P})^\nu}$. The second part is devoted to applications. We construct Hankel invariant distribution spaces. The corresponding test spaces are described in terms of the $\mathscr{S}_\alpha^\beta$-spaces introduced by Gel'fand and Shilov. It turns out that the modified Laguerre polynomials establish an uncountable number of bases for the space of even entire functions in $\mathscr{S}_\mu^\mu$ ($\frac{1}{2}\leqq\mu\leqq 1$). For an even entire function φ we give necessary and sufficient conditions on the coefficients in the Fourier expansion with respect to each basis such that $\varphi\in\mathscr{S}_\mu^\mu$.

AMS(MOS)subject classifications. Primary 46F12, 46F05, 33A65

Introduction. Let X be a separable infinitely dimensional Hilbert space and let $\mathscr{L}$ be a linear operator in X. Then $\mathfrak{D}^\omega(\mathscr{L})$, the analyticity domain of $\mathscr{L}$, consists of all vectors $v\in\bigcap_{n=1}^\infty\mathfrak{D}(\mathscr{L}^n)$ satisfying

$$\exists_{a>0}\exists_{b>0}\forall_{n\in\mathbf{N}}:\|\mathscr{L}^n v\|\leqq n!a^n b.$$

For a positive self-adjoint operator $\mathscr{A}$ in X, Nelson [13] proved that $\mathscr{D}^\omega(\mathscr{A})$ can also be described as

$$\mathfrak{D}^\omega(\mathscr{A})=\bigcup_{t>0}e^{-t\mathscr{A}}(X)=\{e^{-t\mathscr{A}}w\mid w\in X,\ t>0\}.$$

Instead of $\mathfrak{D}^\omega(\mathscr{A})$ we use the notation $\mathscr{S}_{X,\mathscr{A}}$ introduced by de Graaf. The spaces of type $\mathscr{S}_{X,\mathscr{A}}$ are called analyticity spaces. They are nonstrict inductive limits of Hilbert spaces. Together with their strong duals $\mathscr{T}_{X,\mathscr{A}}$ they establish the functional analytic description of the distribution theory in [7].

For each positive constant ν the operator $\mathscr{A}^\nu$ is well defined, positive and self-adjoint in X. So it makes sense to write $\mathscr{S}_{X,\mathscr{A}^\nu}$. The question arises for which perturbations $\mathscr{P}$ on $\mathscr{A}$ there can be found a Hilbert space Y such that $\mathscr{A}+\mathscr{P}$ is a positive self-adjoint operator in Y and $\mathscr{S}_{X,\mathscr{A}^\nu}\subset\mathscr{S}_{Y,(\mathscr{A}+\mathscr{P})^\nu}$. In the paper [1] the case $\nu=1$ has been considered. Also some results concerning analytic dominancy can be found there.

In the second part of this paper we study a class of Hankel invariant test and distribution spaces, and also their relations to the $\mathscr{S}_\alpha^\beta$-spaces of Gel'fand and Shilov [9]. With our papers [2] and [4] we have started this study. There we have shown that the space of even functions in $\mathscr{S}_{1/2}^{1/2}$ remains invariant under the modified Hankel transforms $\mathbf{H}_\alpha$, $\alpha>-1$, defined by

$$(\mathbf{H}_\alpha f)(x)=\int_0^\infty(xy)^{-\alpha}J_\alpha(xy)f(y)y^{2\alpha+1}dy.$$

Moreover, for each $\alpha>-1$ the space of even functions in $\mathscr{S}_{1/2}^{1/2}$ equals the analyticity space $\mathscr{S}_{X_\alpha,\mathscr{A}_\alpha}$ where $X_\alpha=\mathfrak{L}_2((0,\infty),x^{2\alpha+1}dx)$ and $\mathscr{A}_\alpha=-d^2/dx^2+x^2-(2\alpha+1)xd/dx$.

*Received by the editors December 6, 1983.

† Department of Mathematics and Computing Science, Eindhoven University of Technology, Eindhoven, the Netherlands.

The operator $\mathscr{A}_\alpha$ has an orthonormal basis of eigenvectors $(\mathbf{L}_n^{(\alpha)})_{n=0}^\infty$ with eigenvalues $4n+2\alpha+2$. So for each even $f\in\mathscr{S}_{1/2}^{1/2}$ there exists an l_2-sequence $(\omega_n)_{n=0}^\infty$ and $t>0$ such that $f=\sum_{n=0}^\infty \exp(-(4n+2\alpha+2)t)\omega_n\mathbf{L}_n^{(\alpha)}$. Here we prove similar results for the spaces $\mathscr{S}_{X_\alpha,(\mathscr{A}_\alpha)^\nu}$ with $\nu\geqq\frac{1}{2}$ and $\alpha>-1$. It will follow that for all $\alpha,\beta>-1$ and all $\nu\geqq\frac{1}{2}$

$$\mathscr{S}_{X_\alpha,(\mathscr{A}_\alpha)^\nu}=\mathscr{S}_{X_\beta,(\mathscr{A}_\beta)^\nu}.$$

For $\nu\in[\frac{1}{2},1]$ the analyticity space $\mathscr{S}_{X_{-1/2},(\mathscr{A}_{-1/2})^\nu}$ contains just the even functions in $\mathscr{S}_{1/2\nu}^{1/2\nu}$.

1. General theory. Let $\mathscr{A}$ be a positive self-adjoint operator in a Hilbert space X and let $\nu>0$. It makes sense to write $\mathscr{A}^\nu$ and the operator $\mathscr{A}^\nu$ is positive and self-adjoint in X. So the space $\mathscr{S}_{X,\mathscr{A}^\nu}$ is well defined. Its elements are characterized by

LEMMA 1.1. *For each $f\in\mathscr{D}(\mathscr{A}^\infty)\subset X$ the following statements are equivalent:*

(i) $\exists_{a>0}\exists_{b>0}\forall_{k\in\mathbf{N}}:\|\mathscr{A}^k f\|\leqq(k!)^{1/\nu}a^k b$.

(ii) $f\in\mathscr{S}_{X,\mathscr{A}^\nu}$.

Proof. (i)$\Rightarrow$(ii). Let $N\in\mathbf{N}$ and let $\tau>0$. Consider the following estimation

$$(*)\qquad \sum_{k=0}^N \frac{\tau^k}{k!}\|\mathscr{A}^{\nu k}f\|\leqq\sum_{k=0}^N\frac{\tau^k}{k!}\|\mathscr{A}^{-1+\nu k-[\nu k]}\|\|\mathscr{A}^{[\nu k]+1}f\|$$

$$\leqq b_1\sum_{k=0}^N\frac{\tau^k}{k!}(([\nu k]+1)!)^\nu a^k$$

where $b_1=b\sup_{k\in\mathbf{N}\cup\{0\}}(\|\mathscr{A}^{-1+\nu k-[\nu k]}\|)$. The following inequalities are valid:

$$([\nu k]+1)!\leqq([\nu k]+1)([\nu k]+1)^{[\nu k]}\leqq e([\nu k]+1)(\nu k)^{\nu k}.$$

So $([\nu k]+1)!\leqq(e([\nu k]+1))^{1/\nu}(\nu e)^k k!$, and for $\tau<(\nu e a)^{-1}$ the series $(*)$ converges. It implies that $f\in\exp(-\tau\mathscr{A}^\nu)(X)$.

(ii)$\Rightarrow$(i). Suppose $g\in\mathscr{S}_{X,\mathscr{A}^\nu}$. Then there exists $s>0$ and $w\in X$ such that $g=\exp(-s\mathscr{A}^\nu)w$. Let $k\in\mathbf{N}$. Then we estimate as follows

$$\|\mathscr{A}^k f\|\leqq\|\mathscr{A}^k\exp(-s\mathscr{A}^\nu)\|\|w\|=\|w\|\left(\frac{k}{\nu s}\right)^{k/\nu}e^{-k/\nu}$$

$$\leqq\|w\|(1/\nu s)^{k/\nu}\cdot(k!)^{1/\nu}.$$

With $a=(\nu s)^{-1/\nu}$ and $b=\|w\|$ the implication (ii)$\Rightarrow$(i) has been proved. □

Let $\mathscr{L}$ be an unbounded linear operator in X. Then the operators $\mathscr{L}^2,\mathscr{L}^3,\cdots$ are well defined. As a corollary of the previous theorem we get the following.

COROLLARY 1.2. *Let $n\in\mathbf{N}$ and let $f\in\mathscr{D}^\omega(\mathscr{L})$. The following statements are equivalent.*

(i) $\exists_{a>0}\exists_{b>0}\forall_{k\in\mathbf{N}}:\|\mathscr{L}^k f\|\leqq(k!)^{1/n}a^k b$.

(ii) $f\in\mathscr{D}^\omega(\mathscr{L}^n)$.

As mentioned in the introduction we investigate perturbations $\mathscr{P}$ on $\mathscr{A}$ such that $\mathscr{D}^\omega((\mathscr{A}+\mathscr{P})^\nu)\supset\mathscr{S}_{X,\mathscr{A}^\nu}$. For $\nu=1$ the following result has been proved in [1]. Here we consider general $\nu>0$.

THEOREM 1.3. *Let $\mathscr{P}$ be a linear operator in X with $\mathfrak{D}(\mathscr{P}) \supset \mathscr{S}_{X,\mathscr{A}^\nu}$. Suppose the following conditions are satisfied.*

(i) *There exists a Hilbert space Y such that $\exp(-t\mathscr{A}^\nu)$ maps X into Y for all $t>0$.*

(ii) *In addition, $\mathscr{A}+\mathscr{P}$ defined on $\mathscr{S}_{X,\mathscr{A}^\nu}$ is positive and essentially self-adjoint in Y.*

(iii) *There exists an everywhere defined, monotone nonincreasing function φ on $(0,1)$ such that*

$$\forall_{r:0<r<1}: \|\exp(r\mathscr{A}^\nu)\mathscr{P}\mathscr{A}^{-1}\exp(-r\mathscr{A}^\nu)\|_X \leqq \varphi(r).$$

Then $\mathscr{S}_{X,\mathscr{A}^\nu} \subset \mathscr{S}_{Y,(\mathscr{A}+\mathscr{P})^\nu}$.

Proof. We note first that $\mathscr{S}_{X,\mathscr{A}^\nu} = \bigcup_{0<t<1} \exp(-t\mathscr{A}^\nu)(X)$. So let $0<t<1$, and let $0<\tau<t$. Put $s = t-\tau$. We want to estimate the norm of the operator $\exp(\tau\mathscr{A}^\nu)\cdot(\mathscr{A}+\mathscr{P})^k\exp(-t\mathscr{A}^\nu)$ for each $k\in\mathbf{N}$. Therefore we factor as follows

$$\exp(\tau\mathscr{A}^\nu)(\mathscr{A}+\mathscr{P})^k\exp(-t\mathscr{A}^\nu)$$

$$= \prod_{j=0}^{k-1}\left\{\exp\left(\left(\tau+\frac{j}{k}s\right)\mathscr{A}^\nu\right)(\mathscr{I}+\mathscr{P}\mathscr{A}^{-1})\exp\left(-\left(\tau+\frac{j}{k}s\right)\mathscr{A}^\nu\right)\mathscr{A}\exp\left(-\frac{s}{k}\mathscr{A}^\nu\right)\right\}.$$

This factoring yields the estimate

$$\left\|\exp(\tau\mathscr{A}^\nu)(\mathscr{A}+\mathscr{P})^k\exp(-t\mathscr{A}^\nu)\right\|$$

$$\leqq \left\|\mathscr{A}\exp\left(-\frac{s}{k}\mathscr{A}^\nu\right)\right\|^k \prod_{j=0}^{k-1}\left\|\exp\left(\left(\tau+\frac{j}{k}s\right)\mathscr{A}^\nu\right)(\mathscr{I}+\mathscr{P}\mathscr{A}^{-1})\exp\left(-\left(\tau+\frac{j}{k}s\right)\mathscr{A}^\nu\right)\right\|$$

$$\leqq (k!)^{1/\nu}\left(\frac{1}{\nu s}\right)^{k/\nu}\prod_{j=0}^{k-1}\left(1+\varphi\left(\tau+\frac{j}{k}s\right)\right).$$

Since $\varphi(\tau+js/k)\leqq\varphi(\tau)$ for all $j=0,1,\cdots,k-1$, we get

$$\prod_{j=0}^{k-1}\left(1+\varphi\left(\tau+\frac{j}{k}s\right)\right)\leqq(1+\varphi(\tau))^k.$$

Thus we have proved that

$$\forall_{t>0}\forall_{\tau,0<\tau<t}\exists_{a>0}\forall_{k\in\mathbf{N}\cup\{0\}}: \left\|\exp(\tau\mathscr{A}^\nu)(\mathscr{A}+\mathscr{P})^k\exp(-t\mathscr{A}^\nu)\right\|\leqq(k!)^{1/\nu}a^k.$$

Let $t>0$ and let $w\in X$. Set $f=\exp(-t\mathscr{A}^\nu)w$. Then for $0<\tau<t$ fixed there exists $a>0$ such that

$$\|(\mathscr{A}+\mathscr{P})^k f\|_Y \leqq \|\exp(-\tau\mathscr{A}^\nu)\|_{X\to Y}\|\exp(\tau\mathscr{A}^\nu)(\mathscr{A}+\mathscr{P})^k f\|_X$$

$$\leqq \|\exp(-\tau\mathscr{A}^\nu)\|_{X\to Y}\|w\|_X a^k(k!)^{1/\nu}.$$

From Lemma 1.1 it follows that $f\in\mathscr{S}_{Y,(\mathscr{A}+\mathscr{P})^\nu}$. □

Remark. Suppose there exists $k\in\mathbf{N}$ such that the operator $\mathscr{A}^{-k}$ maps X continuously into Y. Then Condition (iii) of Theorem 1.3 is fulfilled because

$$\|\exp(-t\mathscr{A}^\nu)\|_{X\to Y}\leqq\|\mathscr{A}^{-k}\|_{X\to Y}\|\mathscr{A}^k\exp(-t\mathscr{A}^\nu)\|_X.$$

COROLLARY 1.4. *Let $\mathscr{P}$ be an operator in X and let $n\in\mathbf{N}$ with $\mathfrak{D}(\mathscr{P})\supset\mathscr{S}_{X,\mathscr{A}^n}$. Suppose there exists an everywhere defined monotone nonincreasing function φ on $(0,1)$ such that*

$$\forall_{0<r<1}:\left\|\exp(r\mathscr{A}^n)\mathscr{P}\mathscr{A}^{-1}\exp(-r\mathscr{A}^n)\right\|\leqq\varphi(r).$$

Then $\mathscr{S}_{X,\mathscr{A}^n}\subset\mathfrak{D}^\omega((\mathscr{A}+\mathscr{P})^n)$.

Proof. As in the proof of the previous theorem: $\forall_{t>0}\forall_{\tau,0<\tau<t}\exists_{a>0}\forall_{k\in\mathbf{N}}$:

$$\left\|\exp(\tau\mathscr{A}^n)(\mathscr{A}+\mathscr{P})^k\exp(-t\mathscr{A}^n)\right\|\leqq(k!)^{1/n}a^k.$$

So for $f=\exp(-t\mathscr{A}^n)w$, $t>0$, $w\in X$, we get

$$\left\|(\mathscr{A}+\mathscr{P})^k f\right\|_X\leqq\left\|\exp(\tau\mathscr{A}^n)(\mathscr{A}+\mathscr{P})^k\exp(-t\mathscr{A}^n)\right\|\|w\|\leqq(k!)^{1/n}a^k\|w\|.\qquad\square$$

Remark. If $\mathscr{P}$ satisfies the conditions in Corollary 1.4, then $\mathscr{A}^n$ analytically dominates $(\mathscr{A}+\mathscr{P})^n$. (For the terminology, see [6].)

In order to prove the converse statement of Theorem 1.3, i.e.,

$$\mathscr{S}_{Y,(\mathscr{A}+\mathscr{P})^\nu}\subset\mathscr{S}_{X,\mathscr{A}^\nu}$$

we have to interchange the roles of $\mathscr{A}$ and $\mathscr{A}+\mathscr{P}$. Put differently, if we write $\mathscr{B}=\mathscr{A}+\mathscr{P}$ and hence $\mathscr{A}=\mathscr{B}-\mathscr{P}$, then we have to check whether the pair $\mathscr{B}$, $\mathscr{P}$ satisfies the conditions required in Theorem 1.3.

2. Hankel invariant distribution spaces. In our papers [2], [4] on Hankel invariant distribution spaces the following results have been proved.

Let $\mathscr{A}_\gamma$ denote the differential operator $-d^2/dx^2+x^2-(2\gamma+1)/x\,d/dx$ and let X_γ denote the Hilbert space $\mathfrak{L}_2((0,\infty),x^{2\gamma+1}dx)$ where we take $\gamma>-1$. Then for every $\alpha,\beta>-1$ we have shown that

$$\mathscr{S}_{X_\alpha,\mathscr{A}_\alpha}=\mathscr{S}_{X_\beta,\mathscr{A}_\beta}.$$

Moreover, $f\in\mathscr{S}_{X_\gamma,\mathscr{A}_\gamma}$ if and only if f is extendible to an even function in $\mathscr{S}^{1/2}_{1/2}$. Also, it has been proved that the space $\mathscr{S}_{X_\gamma,\mathscr{A}_\gamma}$ remains invariant under the modified Hankel transform $\mathbf{H}_\gamma$ defined by

$$(\mathbf{H}_\gamma f)(x)=\int_0^\infty f(y)(xy)^{-\gamma}J_\gamma(xy)\,y^{2\gamma+1}dy.$$

Here J_γ denotes the Bessel function of the first kind and of order γ. The Hankel transform $\mathbf{H}_\gamma$ extends to a unitary operator on X_γ and $\mathbf{H}_\gamma\mathscr{A}_\gamma=\mathscr{A}_\gamma\mathbf{H}_\gamma$. It follows that for all $\alpha,\beta>-1$, $\mathbf{H}_\alpha$ maps the space $\mathscr{S}_{X_\beta,\mathscr{A}_\beta}$ onto itself. By duality, each $\mathbf{H}_\alpha$ leaves invariant each space of generalized functions $\mathscr{T}_{X_\beta,\mathscr{A}_\beta}$ corresponding to $\mathscr{S}_{X_\beta,\mathscr{A}_\beta}$. The functions $\mathbf{L}_n^{(\gamma)}$ defined by

$$\mathbf{L}_n^{(\gamma)}(x)=\left(\frac{2\Gamma(n+1)}{\Gamma(n+\gamma+1)}\right)^{1/2}e^{-x^2/2}\mathscr{L}_n^{(\gamma)}(x^2),\qquad n\in\mathbf{N}\cup\{0\},\quad x>0$$

establish an orthonormal basis in X_γ and they are the eigenfunctions of the self-adjoint operator $\mathscr{A}_\gamma$ with respective eigenvalues $4n+2\gamma+2$. Here $\mathscr{L}_n^{(\gamma)}$ denotes the nth generalized Laguerre polynomial of order γ. We note that $\mathbf{H}_\gamma\mathbf{L}_n^{(\gamma)}=(-1)^n\mathbf{L}_n^{(\gamma)}$. We recall that for each $\alpha,\beta>-1$ the functions $f\in\mathscr{S}_{X_\alpha,\mathscr{A}_\alpha}$ can be written as $f=\sum_{n=0}^\infty\omega_n\mathbf{L}_n^{(\beta)}$ where $\omega_n=O(e^{-nt})$ for some $t>0$.

With the aid of the theory presented in the first part of this paper we extend the mentioned results and prove that

$$\mathcal{S}_{X_\alpha\{\mathcal{A}_\alpha\}^\nu}=\mathcal{S}_{X_\beta\{\mathcal{A}_\beta\}^\nu}$$

for all $\nu\geqq\frac{1}{2}$ and all $\alpha,\beta>-1$. In addition, we show that for each $\nu\in[\frac{1}{2},1]$ and all $\alpha>-1$ the space $\mathcal{S}_{X_\alpha\{\mathcal{A}_\alpha\}^\nu}$ contains just the even functions of the Gel'fand–Shilov space $\mathcal{S}^{1/2\nu}_{1/2\nu}$. So each even function $f\in\mathcal{S}^{1/2\nu}_{1/2\nu}$ admits Fourier expansions $f=\sum_{n=0}^{\infty}\rho_n^{(\alpha)}\mathbf{L}_n^{(\alpha)}$ with $\rho_n^{(\alpha)}=O(\exp(-n^\nu t))$.

Let $\alpha,\beta>-1$. Then $\mathcal{A}_\alpha$ can be written as

$$\mathcal{A}_\alpha=\mathcal{A}_\beta+2(\alpha-\beta)\mathcal{R}$$

where we put $\mathcal{R}=(1/x)d/dx$. Obviously, $\mathcal{A}_\alpha$ can be obtained from $\mathcal{A}_\beta$ by means of the "perturbation" $2(\alpha-\beta)\mathcal{R}$, and $\mathcal{A}_\beta$ from $\mathcal{A}_\alpha$ by means of $2(\beta-\alpha)\mathcal{R}$. In order to show that $\mathcal{R}$ and hence $c\mathcal{R}$, $c\in\mathbf{C}$, is a perturbation in the sense of Theorem 1.3 we compute the matrix of $\mathcal{R}$ with respect to the orthonormal basis $(\mathbf{L}_n^{(\gamma)})_{n=0}^{\infty}$. To this end, we mention that

$$\mathcal{R}\mathbf{L}_n^{(\gamma)}=-\mathbf{L}_n^{(\gamma+1)}-2\mathbf{L}_{n-1}^{(\gamma+1)}$$

where the relation $d\mathcal{L}_n^{(\gamma)}/dx=-\mathcal{L}_{n-1}^{(\gamma+1)}$ is used.

Now $\mathcal{L}_k^{(\gamma+1)}=\sum_{j=0}^{k}\mathcal{L}_j^{(\gamma)}$ and hence

$$\mathcal{R}\mathbf{L}_n^{(\gamma)}=-\left(\frac{2\Gamma(n+1)}{\Gamma(n+\gamma+1)}\right)^{1/2}\left[\left(\frac{\Gamma(n+\gamma+1)}{\Gamma(n+1)}\right)^{1/2}\mathbf{L}_n^{(\gamma)}+2\sum_{m=0}^{n-1}\left(\frac{\Gamma(m+\gamma+1)}{2\Gamma(m+1)}\right)^{1/2}\mathbf{L}_m^{(\gamma)}\right].$$

Thus we obtain the matrix of $\mathcal{R}$ with respect to the basis $(\mathbf{L}_n^{(\gamma)})_{n=0}^{\infty}$

$$\left(\mathcal{R}\mathbf{L}_k^{(\gamma)},\mathbf{L}_l^{(\gamma)}\right)_\gamma=\begin{cases}-1 & \text{if } l=k,\quad k\in\mathbf{N},\\ 0 & \text{if } l>k,\quad k\in\mathbf{N}\cup\{0\},\\ -2\left(\dfrac{\Gamma(k+1)}{\Gamma(k+\gamma+1)}\dfrac{\Gamma(l+\gamma+1)}{\Gamma(l+1)}\right)^{1/2} & \text{if } 0\leqq l<k,\quad k\in\mathbf{N}.\end{cases}$$

The inequality (cf. [11])

$$n^{1-s}\leqq\frac{\Gamma(n+1)}{\Gamma(n+s)}\leqq(n+1)^{1-s},\qquad 0\leqq s\leqq 1,\quad n\in\mathbf{N}$$

yields

$$\left|\left(\mathcal{R}\mathbf{L}_k^{(\gamma)},\mathbf{L}_l^{(\gamma)}\right)\right|\leqq\begin{cases}2 & \text{if } \gamma\geqq 0,\ 0\leqq l<k,\ k\in\mathbf{N}\cup\{0\},\\ 2k^{-\gamma/2} & \text{if } -1<\gamma<0,\ 0\leqq l<k,\ k\in\mathbf{N}\cup\{0\}.\end{cases}$$

For each $\nu\geqq\frac{1}{2}$, the operator $\exp(r(\mathcal{A}_\gamma)^\nu)\mathcal{R}(\mathcal{A}_\gamma)^{-1}\exp(-r(\mathcal{A}_\gamma)^\nu)$ has to satisfy Condition (iii) of Theorem (1.3). We define the weighted shift operators $\mathcal{W}_{\gamma,\nu}^{(n)}(r)$, $n\in\mathbf{N}\cup\{0\}$,

$$\left(\mathcal{W}_{\gamma,\nu}^{(n)}(r)\mathbf{L}_k^{(\gamma)},\mathbf{L}_l^{(\gamma)}\right)_\gamma=\begin{cases}0 & \text{if } k\neq l+n,\\ \left(\mathcal{R}\mathbf{L}_{l+n}^{(\gamma)},\mathbf{L}_l^{(\gamma)}\right)\dfrac{\exp\left(-r\left(4(l+n)+2\gamma+2\right)\right)^\nu-(4l+2\gamma+2)^\nu}{4(l+n)+2\gamma+2}\end{cases}$$

with norms

(*)

$$\left\|\mathscr{W}_{\gamma,\nu}^{(n)}(r)\right\|_{X_\gamma}=\sup_{l\in\mathbf{N}\cup\{0\}}\left|\left(\mathscr{R}\mathbf{L}_{l+n}^{(\gamma)},\mathbf{L}_l^{(\gamma)}\right)\right|\frac{\exp\left(-r\left(4(l+n)+2\gamma+2\right)\right)^\nu-\left(4l+2\gamma+2\right)^\nu}{4(l+n)+2\gamma+2}.$$

So $\|\mathscr{W}_{\gamma,\nu}^{(0)}(r)\|\leqq 1/(2\gamma+2)$. Now let $n\in\mathbf{N}$. The inequality

$$\left(4(l+n)+2\gamma+2\right)^\nu-\left(4l+2\gamma+2\right)^\nu\geqq(l+n)^{1/2}-l^{1/2}$$

is valid for all $l\in\mathbf{N}\cup\{0\}$ and all $\nu\geqq\frac{1}{2}$. In addition, the matrix elements $|(\mathscr{R}\mathbf{L}_{l+n}^{(\gamma)},\mathbf{L}_l^{(\gamma)})|$ are smaller than $2(l+n)^{-\gamma/2}$ for $-1<\gamma<0$ and smaller than 2 for $\gamma\geqq 0$. If $-1<\gamma\leqq 0$ we therefore get

$$\begin{aligned}\left\|\mathscr{W}_{\gamma,\nu}^{(n)}(r)\right\|&\leqq\sup_{l\in\mathbf{N}\cup\{0\}}\frac{2(l+n)^{-\gamma/2}}{4(l+n)+2\gamma+2}\exp\left(-r\left((l+n)^{1/2}-l^{1/2}\right)\right)\\&\leqq\sup_{l\in\mathbf{N}\cup\{0\}}\left(\frac{1}{2}(l+n)^{-1/2\gamma-1}\exp\left(-\frac{1}{2}rn(l+n)^{-1/2}\right)\right)\\&\leqq\frac{1}{2}\left(1+\frac{1}{2}\gamma\right)^{2+\gamma}\left(\frac{1}{r}\right)^{2+\gamma}\left(\frac{1}{n}\right)^{2+\gamma}\exp(2+\gamma)=:d_1\left(\frac{1}{r}\right)^{2+\gamma}\left(\frac{1}{n}\right)^{2+\gamma}.\end{aligned}$$

Since

$$\exp\left(r\left(\mathscr{A}_\gamma\right)^\nu\right)\mathscr{R}\left(\mathscr{A}_\gamma\right)^{-1}\exp\left(-r\left(\mathscr{A}_\gamma\right)^\nu\right)=\sum_{n=0}^{\infty}\mathscr{W}_{\gamma,\nu}^{(n)}(r)$$

we can use the following straightforward estimate for all $r>0$

$$\begin{aligned}\left\|\exp\left(r\left(\mathscr{A}_\gamma\right)^\nu\right)\mathscr{R}\left(\mathscr{A}_\gamma\right)^{-1}\exp\left(-r\left(\mathscr{A}_\gamma\right)^\nu\right)\right\|&\leqq\sum_{n=0}^{\infty}\left\|\mathscr{W}_{\gamma,\nu}^{(n)}(r)\right\|\\&\leqq\frac{1}{2\gamma+2}+d_1\left(\frac{1}{r}\right)^{2+\gamma}\sum_{n=1}^{\infty}\left(\frac{1}{n}\right)^{2+\gamma}\\&\leqq d_\gamma\left(\frac{1}{r}\right)^{2+\gamma}+\frac{1}{2\gamma+2}\end{aligned}$$

where $d_\gamma=d_1\sum_{n=1}^{\infty}(1/n)^{2+\gamma}$. Summarized we have

LEMMA 2.1. *Let $\gamma>-1$. Then there exist constants $d_\gamma>0$ and $p_\gamma>0$ such that*

$$\forall_{r>0}:\left\|\exp\left(r\left(\mathscr{A}_\gamma\right)^\nu\right)\mathscr{R}\mathscr{A}_\gamma^{-1}\exp\left(-r\left(\mathscr{A}_\gamma\right)^\nu\right)\right\|\leqq d_\gamma\left(\frac{1}{r}\right)^{p_\gamma}+\frac{1}{2\gamma+2}.$$

Proof. For $-1<\gamma\leqq 0$ the assertion has already been proved. For $\gamma>0$ it follows from the matrix expressions for $\mathscr{R}$ that

$$\left\|\exp r\left(\mathscr{A}_\gamma\right)^\nu\mathscr{R}\mathscr{A}_\gamma^{-1}\exp\left(-r\left(\mathscr{A}_\gamma\right)^\nu\right)\right\|\leqq d_0\left(\frac{1}{\gamma}\right)^{p_0}+\frac{1}{2\gamma+2}.$$

□

In addition, we show that given $r>0$, $\gamma,\delta>-1$, the operator $\exp(-r(\mathscr{A}_\gamma)^\nu)$ maps X_γ into X_δ. In [2, p. 17], the following result has been proved

$$\forall_{s\in\mathbf{N}}\exists_{l\in\mathbf{N}}:\left\|\mathscr{Q}^{2s}(\mathscr{A}_\gamma)^{-l}\right\|_\gamma<\infty.$$

Here $\mathscr{Q}$ denotes the multiplication operator in X_γ given by

$$(\mathscr{Q}f)(x)=xf(x).$$

Now let $\delta>-1$ and let $f\in X_\gamma$. Put $s:=[\max\{0,(\delta-\gamma)/2\}]+1$. Then there exists $l_0\in\mathbf{N}$ such that $\|\mathscr{Q}^{2s}\mathscr{A}_\gamma^{-l}\|_\gamma<\infty$ for all $l\geqq l_0$. So we derive

$$(*)\qquad \int_1^\infty\left|\left((\mathscr{A}_\gamma)^{-l}f\right)(x)\right|^2x^{2\delta+1}dx=\int_1^\infty x^{2(\delta-\gamma)}\left|\left((\mathscr{A}_\gamma)^{-l}f\right)(x)\right|^2x^{2\gamma+1}dx$$

$$\leqq\int_1^\infty x^{4s}\left|\left((\mathscr{A}_\gamma)^{-l}f\right)(x)\right|^2x^{2\gamma+1}dx$$

$$\leqq\left\|\mathscr{Q}^{2s}(\mathscr{A}_\gamma)^{-l}\right\|_\gamma^2\|f\|_\gamma^2.$$

Following [12, p. 248], there exists $l_1\in\mathbf{N}$ and $d>0$ such that

$$\max_{x\in[0,1]}\left|\mathbf{L}_k^{(\gamma)}(x)\right|\leqq d(k+1)^{l_1}.$$

For $l>l_1$ it yields

$$(**)\qquad \int_0^1\left|\left((\mathscr{A}_\gamma)^{-2}f\right)(x)\right|^2x^{2\delta+1}dx$$

$$\leqq\left(\max_{x\in[0,1]}\left|\left((\mathscr{A}_\gamma)^{-l}f\right)(x)\right|\right)^2\int_0^1x^{2\delta+1}dx$$

$$\leqq\frac{1}{2\delta+2}\left(\sum_{k=0}^\infty(f,\mathbf{L}_k^{(\gamma)})_\gamma\left(\frac{1}{4k+2\gamma+2}\right)^l\max_{x\in[0,1]}\left|\mathbf{L}_k^{(\gamma)}(x)\right|\right)^2$$

$$\leqq\frac{1}{2\delta+2}\left(d^2\sum_{k=0}^\infty\frac{(k+1)^{2l_1}}{(4k+2\gamma+2)^{2l}}\right)\|f\|_\gamma^2.$$

From $(*)$ and $(**)$ we get

$$\forall_{\gamma>-1}\forall_{\delta>-1}\exists_{l\in\mathbf{N}}\exists_{c>0}\forall_{f\in X_\gamma}:$$

$$\left\|(\mathscr{A}_\gamma)^{-l}f\right\|_\delta^2=\int_0^\infty\left|\left((\mathscr{A}_\gamma)^{-l}f\right)(x)\right|^2x^{2\delta+1}dx\leqq c\|f\|_\gamma^2$$

i.e. $(\mathscr{A}_\gamma)^{-l}$ is a continuous linear operator from X_γ into X_δ.

LEMMA 2.2. *Let* $\gamma>-1$. *Then for every* $r>0$, $\nu>0$ *and* $\delta>-1$ *the operator* $\exp(-r(\mathscr{A}_\gamma)^\nu)$ *is a continuous linear operator from* X_γ *into* X_δ.

Proof. Let $r>0$, $\nu>0$ and let $\delta>-1$. Then there exists $l\in\mathbf{N}$ such that $(\mathscr{A}_\gamma)^{-l}$ is a continuous linear mapping from X_γ into X_δ. Hence $\exp(-r(A_\gamma)^\nu)=(\mathscr{A}_\gamma)^{-l}\{(\mathscr{A}_\gamma)^l\exp(-r(\mathscr{A}_\gamma)^\nu)\}$ is also a continuous linear mapping from X_γ into X_δ. □

Lemmas 2.1 and 2.2 yield the following important result.

THEOREM 2.3. *Let $\alpha, \beta > -1$. Then for every $\nu \geqq \frac{1}{2}$*

$$\mathcal{S}_{X_\alpha,(\mathcal{A}_\alpha)^\nu} = \mathcal{S}_{X_\beta,(\mathcal{A}_\beta)^\nu}.$$

Proof. Let $\nu \geqq \frac{1}{2}$. We have shown that:
$-\exp(-t(\mathcal{A}_\alpha)^\nu)$, $t>0$, maps X_α continuously into X_β.
$-\mathcal{D}(\mathcal{R}) \supset \mathcal{S}_{S_\alpha,(\mathcal{A}_\alpha)^\nu}$, and $\mathcal{A}_\beta = \mathcal{A}_\alpha + 2(\alpha-\beta)\mathcal{R}$ is positive and self-adjoint in X_β.
$-$There exist constants $d_\alpha, p_\alpha > 0$ such that for all $r>0$

$$\left\| \exp\left(r(\mathcal{A}_\alpha)^\nu\right) \mathcal{R}(\mathcal{A}_\alpha)^{-1} \exp\left(-r(\mathcal{A}_\alpha)^\nu\right) \right\|_\alpha \leqq d_\alpha \left(\frac{1}{r}\right)^{p_\alpha} + \frac{1}{2\alpha+2}.$$

So by Theorem 1.3, $S_{X_\alpha,(\mathcal{A}_\alpha)^\nu} \subset \mathcal{S}_{X_\beta,(\mathcal{A}_\beta)^\nu}$. Interchanging α and β we get the wanted result. □

Let $\alpha > -1$. Since $\mathbf{H}_\alpha \mathcal{A}_\alpha = \mathcal{A}_\alpha \mathbf{H}_\alpha$, also $\mathbf{H}_\alpha (\mathcal{A}_\alpha)^\nu = (\mathcal{A}_\alpha)^\nu \mathbf{H}_\alpha$. So the Hankel transform $\mathbf{H}_\alpha$ is a continuous bijection on the space $\mathcal{S}_{X_\alpha,(\mathcal{A}_\alpha)^\nu}$, $\nu \geqq \frac{1}{2}$, and hence on the spaces $\mathcal{S}_{X_\beta,(\mathcal{A}_\beta)^\nu}$, $\nu \geqq \frac{1}{2}$, $\beta > -1$. By duality each transform $\mathbf{H}_\alpha$ leaves invariant the spaces of generalized functions $\mathcal{T}_{X_\beta,(\mathcal{A}_\beta)^\nu}$. For $\alpha = -\frac{1}{2}$ we get $X_{-1/2} = \mathfrak{L}_2((0,\infty))$ and $\mathcal{A}_{-1/2} = -(d^2/dx^2) + x^2$. The functions $\mathbf{L}_k^{(-1/2)}$ are the even Hermite functions. With the aid of the papers [8] and [10] the following characterization of the spaces $\mathcal{S}_{X_{-1/2},(\mathcal{A}_{-1/2})^\nu}$, $\nu \in [\frac{1}{2}, 1]$, can be obtained,

$$f \in \mathcal{S}_{X_{-1/2},(\mathcal{A}_{-1/2})^\nu} : \Leftrightarrow f \text{ is extendible to an even function in the space } \mathcal{S}_{1/2\nu}^{1/2\nu}.$$

The spaces $\mathcal{S}_p^q$, $p+q \geqq 1$, $p,q \geqq 0$, are introduced by Gel'fand and Shilov in [9]. In this connection we note that in our paper [5] we have proved that the spaces $\mathcal{S}_{1/k+1}^{k/k+1}$ are analyticity spaces; explicitly

$$\mathcal{S}_{1/k+1}^{k/k+1} = \mathcal{S}_{\mathfrak{L}_2(\mathbf{R}),\mathcal{B}_k} \quad \text{with } \mathcal{B}_k = \left(-\frac{d^2}{dx^2} + x^{2k} \right)^{(k+1)/2k}.$$

Relevant for the present paper are the spaces $\mathcal{S}_\mu^\mu$, $\frac{1}{2} \leqq \mu \leqq 1$. We have

$\varphi \in \mathcal{S}_\mu^\mu$, $\frac{1}{2} \leqq \mu \leqq 1$ if and only if φ is an entire function satisfying $\exists_{A,B,C>0} : |\varphi(x+iy)| \leqq C \exp(-A|x|^{1/\mu} + B|y|^{1/1-\mu})$

and

$\varphi \in \mathcal{S}_1^1$ if and only if φ is analytic on a strip about the real axis say of width $r>0$ and satisfying $\exists_{A,C>0} : \sup_{|y|<r} |\varphi(x+iy)| \leqq C \exp(-A|x|)$.

Now Theorem 2.3 leads to the following important results.

COROLLARY 2.4. *Let $\alpha > -1$ and let $\nu \in [\frac{1}{2}, 1]$. Then $f \in \mathcal{S}_{X_\alpha,(\mathcal{A}_\alpha)^\nu}$ if and only if f is extendible to an even function in the space $\mathcal{S}_{1/2\nu}^{1/2\nu}$.*

COROLLARY 2.5. *Let $f \in \mathcal{S}_{1/2\nu}^{1/2\nu}$ be even, with $\nu \in [\frac{1}{2}, 1]$. Then for each $\gamma > -1$, there exists an l_2-sequence $(\omega_n^{(\gamma)})_{n=0}^\infty$ and $t>0$ such that $f = \sum_{n=0}^\infty \exp(-n^\nu t)\omega_n^{(\gamma)} \mathbf{L}_n^{(\gamma)}$ where the series converges pointwise.*

Appendix. The set of so-called entire vectors for a positive self-adjoint operator $\mathscr{A}$ in a Hilbert space X is equal to

$$\mathfrak{D}^\infty(e^{\mathscr{A}}) = \bigcap_{t>0} e^{-t\mathscr{A}}(X).$$

In [3], van Eijndhoven has used the Fréchet space $\mathfrak{D}^\infty(e^{\mathscr{A}})$ as the test space in a theory of generalized functions which is a kind of reverse of the theory in [7]. The space $\mathfrak{D}^\infty(e^{\mathscr{A}})$ is denoted by $\tau(X,\mathscr{A})$ and it may be called the entireness space. To our opinion the well-known theory of tempered distributions is considerably generalized in [3]. (Put $\mathscr{A}=\log(-d^2/dx^2+x^2+1)$. Then $\tau(\mathfrak{L}_2(\mathbf{R}), \mathscr{A})$ is the space $\mathscr{S}(\mathbf{R})$ of functions of rapid decrease.)

Similar to Theorem 1.3 we prove

THEOREM A.1. *Let $\mathscr{P}$ be a linear operator in X with $\mathfrak{D}(\mathscr{P})\supset \exp(-\sigma\mathscr{A}^\nu)(X)$ for some $\sigma>0$ sufficiently large. Suppose the following conditions are satisfied.*

(i) *There exists a Hilbert space Y such that $\exp(-t\mathscr{A}^\nu)$ maps X into Y for all $t>0$.*

(ii) *Also, $\mathscr{A}+\mathscr{P}$ defined on $\exp(-\sigma\mathscr{A}^\nu)(X)$ is a positive essentially self-adjoint operator in Y.*

(iii) *There exist positive constants $r_0\geqq 1$, $d>0$ and $0\leqq q<1/\nu$ such that for all $r>r_0$*

$$\|\exp(r\mathscr{A}^\nu)\mathscr{P}\mathscr{A}^{-1}\exp(-r\mathscr{A}^\nu)\|_X < dr^q.$$

Then $\tau(X,\mathscr{A}^\nu)\subset\tau(Y,(\mathscr{A}+\mathscr{P})^\nu)$.

Proof. Since $\tau(X,\mathscr{A}^\nu)=\bigcap_{t>r_0}\exp(-t\mathscr{A}^\nu)(X)$, we consider $t>r_0$ only. Let $0<\tau<1$ with $s=t-\tau>1$. The factoring used in Theorem 1.3 yields the following estimate

$$\|\exp(\tau\mathscr{A}^\nu)(\mathscr{A}+\mathscr{P})^k\exp(-t\mathscr{A}^\nu)\| \leqq k!\left(\frac{1}{\nu s}\right)^{k/\nu}\prod_{j=0}^{k-1}(1+d(\tau+js/k)^q).$$

Put $b_\tau=1+d\tau^q$. Then

$$\prod_{j=0}^{k-1}(1+d(\tau+js/k)^q)\leqq b_\tau\prod_{j=1}^{k-1}(1+d)\left(\frac{k+js}{k}\right)^q\leqq b_\tau(1+d)^k 2^{qk}s^{qk}.$$

Set $a=(1+d)2^q(\frac{1}{\nu})^{1/\nu}$. Then

$$\|\exp(\tau\mathscr{A}^\nu)(\mathscr{A}+\mathscr{P})^k\exp(-t\mathscr{A}^\nu)\|_X\leqq(k!)^{1/\nu}\left(\frac{1}{s}\right)^{(-q+1/\nu)k}a^k b_\tau.$$

For $f\in\exp(-t\mathscr{A}^\nu)(X)$ it yields

$$\begin{aligned}\|(\mathscr{A}+\mathscr{P})^k f\|_Y &\leqq \|\exp(-\tau A^\nu)\|_{X\to Y}\|\exp(\tau\mathscr{A}^\nu)(\mathscr{A}+\mathscr{P})^k\exp(-t\mathscr{A}^\nu)\|_X\|\exp(t\mathscr{A}^\nu)f\| \\ &\leqq (k!)^{1/\nu}\left(a\cdot\left(\frac{1}{s}\right)^{1/\nu-q}\right)^k b_\tau\|\exp(-\tau\mathscr{A}^\nu)\|_{X\to Y}\|\exp(t\mathscr{A}^\nu)f\|_X.\end{aligned}$$

Thus we find that $f\in\exp(-r(\mathscr{A}+\mathscr{P})^\nu)(Y)$ for all $r<(1/\nu ae)s^{-q+1/\nu}$. Now put $r(t)=(1/(\nu ae+1))s^{-q+1/\nu}$ with $s=t+1/t-1$ for instance. Then we get

$$\begin{aligned}\tau(X,\mathscr{A}^\nu) &= \bigcap_{t>r_0}(\exp(-t\mathscr{A}^\nu)(X))\subset\bigcap_{t>r_0}\left(\exp(-r(t)(\mathscr{A}+\mathscr{P})^\nu)(Y)\right) \\ &= \bigcap_{r>0}\left(\exp(-r(\mathscr{A}+\mathscr{P})^\nu)(Y)\right)\Big) = \tau(Y,(\mathscr{A}+\mathscr{P})^\nu).\end{aligned}$$

□

It is not hard to see that the spaces $\tau(X_\alpha, (\mathscr{A}_\alpha)^\nu)$, $\alpha > -1$, are Hankel invariant, and hence their strong duals $\sigma(X_\alpha, (\mathscr{A}_\alpha)^\nu)$. The previous theorem and the Lemmas 2.1 and 2.2 lead to the following classification.

THEOREM A.2. *Let* $\alpha, \beta > -1$ *and let* $\nu \geqq \frac{1}{2}$. *Then*

$$\tau\left(X_\alpha, (\mathscr{A}_\alpha)^\nu\right) = \tau\left(X_\beta, (\mathscr{A}_\beta)^\nu\right).$$

By [2] *and* [8] *we obtain the following characterizations*

$f \in \tau(X_{-1/2}, \mathscr{A}_{-1/2})$ *iff* f *is extendible to an even entire function for which* $\forall_{0<a<1} \exists_{C>0} \forall_{x+iy \in C}: |f(x+iy)| \leqq C \exp(-\frac{1}{2}ax^2 + \frac{1}{2a}y^2)$

and

$f \in \tau(X_{-1/2}, (\mathscr{A}_{-1/2})^{1/2})$ *iff* f *is extendible to an even entire function for which* $\forall_{r>0}: \sup_{|y|<r, -\infty<x<\infty} e^{r|x|} |f(x+iy)| < \infty$.

Finally, Theorem A.2 gives the characterization in classical analytic terms of the elements in each $\tau(X_\alpha, \mathscr{A}_\alpha)$, respectively $\tau(X_\alpha, (\mathscr{A}_\alpha)^{1/2})$, $\alpha > -1$.

REFERENCES

[1] S. J. L. VAN EIJNDHOVEN, *Invariance of the analyticity domain of self-adjoint operators subjected to perturbations*, preprint, 1982.

[2] _____, *On Hankel invariant distribution spaces*, EUT-Report 82-WSK-01, Eindhoven Univ. of Technology, Eindhoven, the Netherlands, 1982.

[3] _____, *A theory of generalized functions based on one-parameter groups of unbounded self-adjoint operators*, TH-Report 81-WSK-03, Eindhoven Univ. of Technology, Eindhoven, the Netherlands, 1981.

[4] S. J. L. VAN EIJNDHOVEN AND J. DE GRAAF, *Some results on Hankel invariant distribution spaces*, Proc. Koninklijke Nederlands Akademie van Wetenschappen, A(86)1, 1983.

[5] S. J. L. VAN EIJNDHOVEN, J. DE GRAAF AND R. S. PATHAK, *A characterization of the spaces* $\mathscr{S}_1^{k/k+1}$ *by means of holomorphic semigroups*, this Journal, 14 (1983), pp. 1180–1187.

[6] W. G. FARIS, *Self-Adjoint Operators*, Lecture Notes in Mathematics 433, Springer, Berlin, 1974.

[7] J. DE GRAAF, *A theory of generalized functions based on holomorphic semigroups*, (three papers), Proc. Koninklijke Nederlandse Akademie van Wetenschappen. A 86(4), 1983, pp. 407–420; A 87(2), 1984, pp. 155–171; A 87(2), 1984, pp. 173–187.

[8] R. GOODMAN, *Analytic and entire vectors for representation of Lie groups*, Trans. Amer. Math. Soc., 143 (1969), pp. 5–76.

[9] I. M. GEL'FAND AND G. E. SHILOV, *Generalized Functions*, *Vol.* II, Academic Press, New York, 1968.

[10] ZHANG GONG-ZHING, *Theory of distributions of S-type and expansions*, Chinese Math., 4(2) (1963), pp. 211–221.

[11] D. S. MITRINOVIC, *Analytic Inequalities*, first ed., Springer, Berlin, 1970.

[12] W. MAGNUS, F. OBERHETTINGER AND R. P. SONI, *Formulas and Theorems for the Special Functions of Mathematical Physics*, third ed., Springer, Berlin, 1966.

[13] E. NELSON, *Analytic vectors*, Ann. Math., 70 (1959), pp. 572–615.

SIAM J. MATH. ANAL.
Vol. 17, No. 3, May 1986

LIMIT CYCLES IN THE JOSEPHSON EQUATION*

JAN A. SANDERS† AND RICHARD CUSHMAN‡

Abstract. Using techniques from bifurcation theory, we find the bifurcation diagram and corresponding phase portraits on TS^1 of the Josephson equation: $\dot{\phi}=y$, $\dot{y}=-\sin\phi+\varepsilon(a-(1+\gamma\cos\phi)y)$.

AMS(MOS) subject classifications. Primary 58F21; secondary 34C05

Key words. Josephson equation, averaging method, Picard–Fuchs equation, limit cycle, saddle connection, homoclinic loop

1. Introduction. Mathematically, a single point contact Josephson junction is described by the nonlinear second order differential equation

$$\beta\frac{d^2\phi}{d\tau^2}+(1+\gamma\cos\phi)\frac{d\phi}{d\tau}+\sin\phi=\alpha, \qquad \phi\in S^1 \quad \alpha,\beta,\gamma\in\mathbb{R} \tag{1.1}$$

which we call the Josephson equation. In many applications [14], [19] point contact Josephson junctions are used as precision voltage sources. Here the voltage of a solution of (1.1) is the time average of $d\phi/d\tau$. Unfortunately the output voltage of a Josephson junction is very small. In order to physically detect it one first drives the Josephson junction with a high frequency sine wave of low amplitude and then one measures the voltage of a phased locked solution. Mathematically, one is looking at the time average of $d\phi/d\tau$ on a periodic solution which lies on an invariant torus of the driven equation. Because the amplitude of the driving is small, one expects that the invariant tori come from limit cycles of the Josephson equation. Therefore one needs to know the entire phase portrait of (1.1).

We shall study (1.1) when β is large and α is small. More specifically, let $\varepsilon=\beta^{-1/2}$ be a small positive parameter. Then set $\alpha=\varepsilon a$ and $\tau=\varepsilon t$. With $\dot{}=d/dt$ we write (1.1) as

$$X_{a,\gamma}: \quad \begin{aligned}\dot{\phi}&=y,\\ \dot{y}&=-\sin\phi+\varepsilon[a-(1+\gamma\cos\phi)y].\end{aligned} \tag{1.2}$$

Holding ε fixed, $X_{a,\gamma}$ is a two-parameter family of vectorfields on the cylinder TS^1, which we will study by the averaging method.

We now derive the averaged equation. When $\varepsilon=0$ (1.2) is the Hamiltonian vectorfield X_H describing the mathematical pendulum where the Hamiltonian function is

$$H(\phi,y)=\frac{1}{2}y^2-\cos\phi. \tag{1.3}$$

Instead of the variables (ϕ,y) we will use the variables (ϕ,h), where h is defined as

$$h=\frac{1}{2}y^2-\cos\phi.$$

*Received by the editors September 28, 1984, and in revised form March 1, 1985.

†Department of Mathematics and Computer Science, Free University, 1007 MC Amsterdam, the Netherlands.

‡Department of Mathematics and Computer Science, Free University, 1007 MC Amsterdam, the Netherlands. Present address, Mathematics, Institute, University of Utrecht, Budapestlaan 6, De Uithof, Utrecht, the Netherlands.

Differentiating (1.3) with respect to ϕ and using (1.2) gives

$$\frac{dh}{d\phi}=\varepsilon[a-(1+\gamma\cos\phi)y(\phi,h)] \tag{1.4}$$

where $y(\phi,h)=\pm(\sqrt{2}(h+\cos\phi))^{1/2}$. Averaging (1.4) over a compact connected component Γ_h of the level set $H^{-1}(h)$ (which is a periodic solution of X_h except when $h=1$) leads to the averaged equation

$$\frac{\overline{dh}}{d\phi}=\varepsilon\left[a\int_{\Gamma_h}d\phi-\int_{\Gamma_h}y\,d\phi-\gamma\int_{\Gamma_h}y\cos\phi\,d\phi\right]. \tag{1.5}$$

Nondegenerate zeros of the right-hand side of (1.5) correspond to limit cycles of $X_{a,\gamma}$.

In earlier publications [3], [17], the averaged equation was studied by expressing the integrals

$$\mathscr{A}(h)=\int_{\Gamma_h}d\phi,\quad \mathscr{B}(h)=\int_{\Gamma_h}y\,d\phi,\quad \mathscr{C}(h)=\int_{\Gamma_h}y\cos\phi\,d\phi, \tag{1.6}$$

in terms of the complete elliptic integrals E and K. Because of the complicated implicit relationship between the elliptic modulus parameter k and the energy parameter h, this special function approach led to meager incomplete results. The technique we use to study the averaged equation is to find the Picard–Fuchs equation satisfied by $\mathscr{A}$, $\mathscr{B}$ and $\mathscr{C}$ and then to analyze the solutions of the resulting Riccati equations. This is the approach used in the study of codimension two bifurcations of planar vectorfields [4]–[9], [11]–[13] and is quite widely applicable. For instance, for all perturbations of the mathematical pendulum by a vectorfield on TS^1 whose components are polynomical in y, $\cos\phi$, and $\sin\phi$, there is a systematic way of finding the Picard–Fuchs equation for the integrals appearing in the averaged equation. However for the analysis of the resulting Riccati equations there are no known general techniques. Here we must proceed on a case by case basis, although one can obtain upper estimates on the number of solutions in terms of the degree of the polynomial [18].

We return to studying the zeros of the averaged equation (1.5). First, we remark that there are *three* distinct families of closed cycles Γ_h on TS^1 (see Fig. 1):

i. Γ_h^0, when $-1<h<1$. The level set $H^{-1}(h)$ is smooth, connected, compact, and contractible to a point.

ii. Γ_h^+, when $h>1$ and $y>0$. Γ_h^+ is the component of the level set $H^{-1}(h)$ given by the graph of the function $y=\sqrt{2(h+\cos\phi)}$; Γ_h^+ is not contractible to a point (it winds around the cylinder).

iii. Γ_h^-, when $h>1$ and $y<0$. Γ_h^- is as in (ii), except that $y=-\sqrt{2(h+\cos\phi)}$.

From here on we use the superscripts 0 and $\pm$ on $\mathscr{A}$, $\mathscr{B}$ and $\mathscr{C}$ to denote which Γ_h-family is being used. For a fixed value of the parameter a, those values of γ which give rise to zeros of $\overline{dh}/d\phi$ in (1.5) are exactly the values of the function

$$\eta(h)=a\frac{\mathscr{A}(h)}{\mathscr{C}(h)}-\xi(h), \tag{1.7}$$

where

$$\xi(h)=\frac{\mathscr{B}(h)}{\mathscr{C}(h)}. \tag{1.8}$$

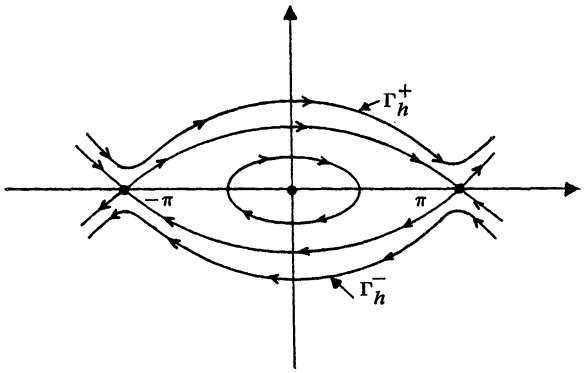


FIG. 1. *Level sets of H on* TS^1.

The main result of this paper is the following.

THEOREM. *For* $h \in (-1,1)$, $\eta^0 = -\xi^0$ *is a strictly monotonic decreasing function with range* $(-3,-1)$; *moreover* η^0 *does not depend on the parameter a.*

Let $a^{\pm} = \pm 16/3\pi$. *For each* $a < a^+$, η^+ *is a strictly monotonic decreasing function of h on* $(1,\infty)$ *with range* $(-\infty, 3\pi a/4 - 3)$; *while for each* $a > a^+$, *when* $h > 1$, η^+ *has a unique maximum* $\tilde{\eta}^+(a)$ *and has range* $(-\infty, \tilde{\eta}^+(a))$. *For each* $a > a^-$, η^- *is strictly monotonic decreasing function on* $(1,\infty)$ *with range* $(-\infty, 3\pi a/4 - 3)$; *while for each* $a < a^-$, *when* $h > 1$, η^- *has a unique maximum* $\tilde{\eta}^-(a)$ *and has range* $(-\infty, \tilde{\eta}^-(a))$.

Figure 2 gives the a-γ plane bifurcation diagram for $X_{a,\gamma}$, while Fig. 3 depicts the phase portraits of $X_{a,\gamma}$, along two curves in the a-γ plane. In §2 we will show that Figs. 2 and 3 follow from the above theorem. In §3 and §4 we give somewhat technical proofs of the key properties of the functions ξ and η, which culminate in the proof of the theorem.

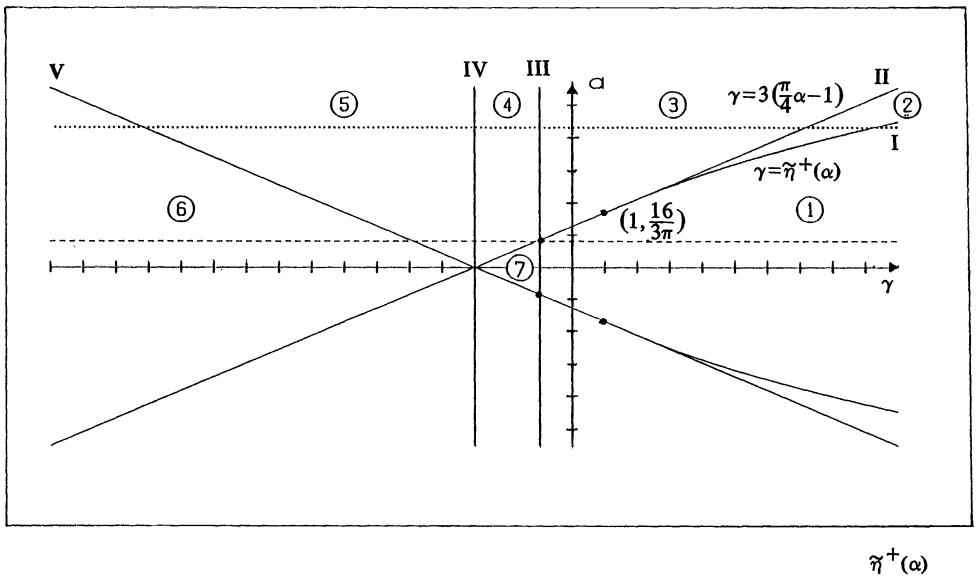
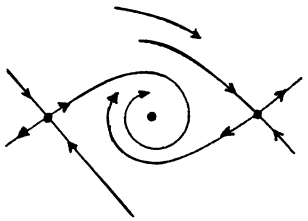
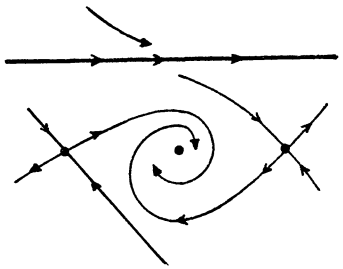


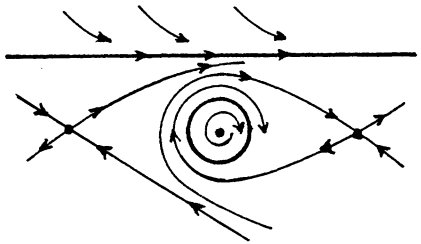
FIG. 2. *Bifurcation set for* $X_{a,\gamma}$.



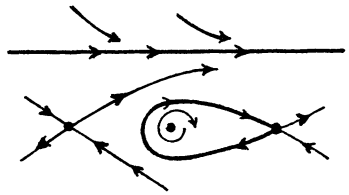
1 $|\alpha| < 8/3\pi$. *An attracting sink and saddle point.*



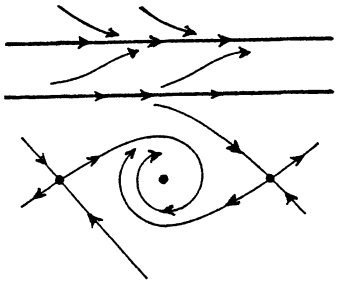
I. *Appearance of a semistable winding limit cycle.*



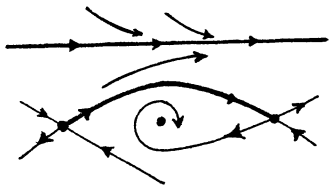
4 *Growth of stable Hopf limit cycle.*



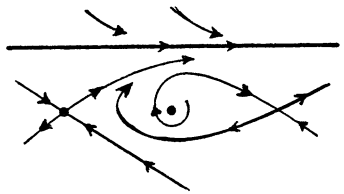
IV. *Death of stable Hopf limit cycle in formation of homoclinic loop.*



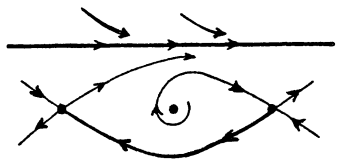
2 *Splitting of semistable winding limit cycle into stable-unstable pair of winding limit cycles.*



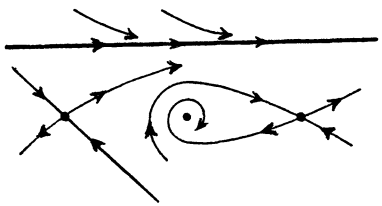
II. $\gamma > 1$. *Formation of upper saddle connection by unstable winding limit cycle.*



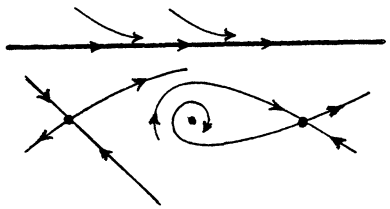
5 *Dissolution of homoclinic loop.*



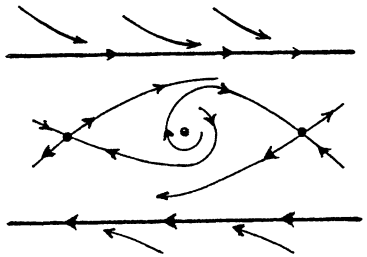
V. *Formation of lower saddle connection.*



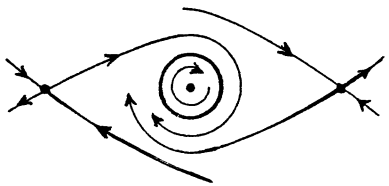
3 *Dissolution of upper saddle connection.*



III. $|\alpha| > 8/3\pi$. *Birth of stable Hopf limit cycle.*

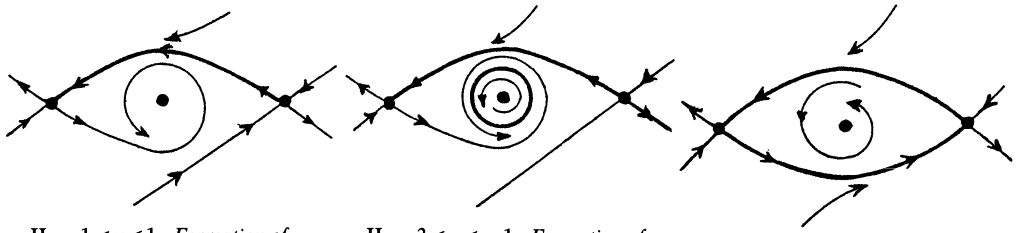


6 *Dissolution of lower saddle connection and formation of lower winding limit cycle.*



7 *Growth of stable Hopf limit cycle.*

FIG. 3. *Phase portraits of* $X_{a,\gamma}$.



II. $-1 \leqq \gamma \leqq 1$. *Formation of upper saddle connection with no other limit cycle present.*

II. $-3 < \gamma < -1$. *Formation of upper saddle connection in the presence of a stable Hopf limit cycle.*

II. $\gamma = -3$, $a = 0$ *Formation of a double saddle connection.*

FIG. 3 (*cont.*) *Phase portraits of* $X_{a,\gamma}$.

2. Phase portraits. Consider Figs. 2 and 3. Following the dotted line in Fig. 2 from the right to the left we go through numbered regions and curves with Roman numerals corresponding with phase portraits in Fig 2. We shall now see why the phase portraits follow from our knowledge of η. Before beginning our discussion we would like to warn the ambitious reader who tries to fill in the details omitted in our severely condensed arguments that the Poincaré–Bendixson theorem [10] does not help in finding the α and ω limit sets of orbits of $X_{a,\gamma}$ because TS^1 is not a compact and hence the orbits may run off to infinity. Also one should note that the divergence criterion holds only on simply connected regions.

The sections correspond to the number of the region in Figs. 2 and 3. Note that Fig. 3 is symmetric about the γ axis since (1.1) is the same when (a, ϕ) is replaced by $(-a, -\phi)$. Also the phase portraits are the same when a is replaced by $-a$. Thus we suppose $a \geqq 0$.

1. First we show that there is a $y_0 > 0$ such that $\mathcal{D} = \{(\phi, y) \in TS^1 \,|\, |y| \leqq y_0\}$ is an attracting domain for $X_{a,\gamma}$. From the averaging theorem it follows that an integral curve on $X_{a,\gamma}$ starting at p lies within an ε tube of an integral curve of X_H starting at p for all $t \in [0, C\varepsilon^{-1}]$. Choose y_0 so large that every integral curve of X_H which starts outside $\mathcal{D}$ has a period less than $\frac{1}{2}C\varepsilon^{-1}$. Therefore any vertical line $l = \{(\phi, y) \in TS^1 \,|\, \phi = \phi_0\}$ is a global cross-section for $X_{a,\gamma}$ outside $\mathcal{D}$. Again, using the averaging theorem, it follows that outside $\mathcal{D}$ the change in H along an integral curve of $X_{a,\gamma}$ between successive crossings of l is equal to $\overline{dh/d\phi}$ plus terms of order ε^2. Since (a, γ) lies in region 1,

$$\eta^{0,+}(h) > \gamma \tag{2.1}$$

for all $h > -1$. From (4.1), the definition of η (1.7) and the fact that $C^{0,+}(h) > 0$ (3.4 & 3.29), it follows that $\overline{dh/d\phi} < 0$. Therefore, on each integral curve of $X_{a,\gamma}$ starting outside $\mathcal{D}$, H is strictly decreasing. Consequently, $\mathcal{D}$ is a compact attracting domain for $X_{a,\gamma}$.

The original system (1.2) has stationary points given by

$$\sin\phi = \varepsilon a, \qquad y = 0.$$

Thus ϕ is approximately 0 or π. In what follows we shall denote the two equilibrium points by $(0, 0)$ and $(\pi, 0)$ to keep the notation simple. The linearization of (1.2) at these stationary points is approximately

$$\begin{pmatrix} 0 & 1 \\ \mp 1 & -\varepsilon(1 \pm \gamma) \end{pmatrix} \tag{2.2}$$

where the + corresponds with $(0,0)$ and the $-$ with $(\pi,0)$. The point $(\pi,0)$ is a saddle point. At $(0,0)$ the trace of (2.2) is $-\varepsilon(1+\gamma)$. Thus for $\gamma>-1$ the point $(0,0)$ is attracting.

We now finish the discussion of the phase portrait of $X_{a,\gamma}$. Since (2.1) holds, $X_{a,\gamma}$ has no limit cycles or saddle connections in $\mathscr{D}$. Therefore by the Poincaré–Bendixson theorem, the ω-limit set of every integral curve of $X_{a,\gamma}$ in $\mathscr{D}$ is an equilibrium point. Since $X_{a,\gamma}$ has no saddle connections, the ω-limit of the unstable manifold of $(\pi,0)$ is $(0,0)$. Therefore $(0,0)$ is a global attractor for $X_{a,\gamma}$ except for the stable manifold of $(\pi,0)$ whose α-limit set runs off to infinity.

2. On I: $\gamma=\eta^+(a)$, we witness a saddle node type bifurcation of a stable and unstable winding limit cycle. In region 2, the unstable winding limit cycle divides the attraction domains of the stable winding limit cycle and the two stationary points.

3. On II: $\gamma=3(\pi a/4-1)$, $\gamma>1$, the unstable winding limit cycle forms a noncontractible homoclinic saddle connection. This can also be deduced from the vanishing of the Melnikov function for those readers who prefer to use techniques other than averaging. If, after crossing II, the ω-limit set of both branches of the unstable manifold again would be the point $(0,0)$, then there is no way for the α-limit set of the stable manifold to exist. Therefore, since the Melnikov function changes sign at the bifurcation, the ω-limit set consists of the point $(0,0)$ and the attracting winding limit cycle. The domains of attraction are separated by the stable manifold of the saddle point. The existence of contractible limit cycles is ruled out by the fact that the range of η^0 is $[-3,-1]$ and, for $|\gamma|<1$, by the negativity of the divergence of $X_{a,\gamma}$.

4. Crossing III: $\gamma=-1$, a Hopf bifurcation takes place. This gives rise to the existence of a stable contractible limit cycle for $\gamma\in[-3,-1]$, the range of η^0. The Hopf limit cycle takes over the role of the point attractor as described in part 3. The part of the unstable manifold that is approaching the contractible limit cycle is characterized by the fact that it intersects the line segment $y=0$, $\pi<\phi<2\pi$, i.e. it intersects the ϕ-axis going from the positive ($y>0$) half to the negative half of the cylinder. Due to the growth of the contractible limit cycle, as reflected in the monotonicity of η^0, this intersection takes place closer and closer to the saddle point, until, at IV: $\gamma=-3$, a homoclinic (contractible) saddle connection is formed. The existence of this homoclinic connection follows from a continuity argument relating the behavior of the unstable manifold for $\gamma>-3$ to that for $-3-3\pi a/4<\gamma<-3$ (see the next section).

5. For $\gamma<-3$, the only attractor is the winding limit cycle. The point $(0,0)$ stays unstable. There is one orbit going out of $(0,0)$ to the saddle point. Both branches of the unstable manifold of the saddle point are attracted to the winding limit cycle. The only other possibilities would be:

i. The unstable manifold forms a noncontractible homoclinic saddle connection. This is ruled out by the Melnikov function.

ii. It forms a contractible homoclinic connection (as for $\gamma=-3$). If we average the original equation along this connection, we find that $\overline{dh}/d\phi$ is strictly positive. This is a contradiction. A branch of unstable manifold of the saddle point cannot cross the line segment S: $y=0$ from $(0,0)$ to $(\pi,0)$, since if it did this would imply the existence of a stable contractible limit cycle (because the point $(0,0)$ is unstable). Therefore this branch has to cross $y=0$ between $(-\pi,0)$ and $(0,0)$. Thereafter it approaches the winding limit cycle. Since the branch of the unstable manifold with initial negative velocity crosses the line segment S when $\gamma>-3$, it follows by continuity that for γ near -3 this branch of the unstable manifold goes into the saddle point and thus forms a contractible homoclinic connection.

6. At V: $\gamma = -3 - 3\pi a/4$, γ is in the range of η^-. One branch of the unstable manifold of the saddle point forms a noncontractible homoclinic saddle connection, which, for smaller γ, splits off as a stable winding limit cycle. The unstable manifold now has both winding limit cycles as its ω-limit set, while the stable manifold has $(0,0)$ as its α-limit set. This concludes our discussion of the phase portraits along the dotted line in Fig. 3.

Along the dashed line in Fig. 3, the main difference is that the stable limit cycle is not born from a saddle node type bifurcation, but from a noncontractible homoclinic saddle connection, as in the transition from 5 to 6. Since the techniques needed to analyze all other phase portraits and their bifurcations are essentially the same as we have used thus far, we shall not give any more details.

3. Properties of the function ξ. In this section we show that the function ξ has the following properties:

i. For $h \in (-1,1)$, $\eta^0 = -\xi^0$; for $h \in (1,\infty)$, $\xi^- = \xi^+$.

ii. $\xi = \xi^{0,\pm}$ satisfies the Riccati equation

$$2(1-h^2)\frac{d\xi}{dh} = 3 + 2h\xi - \xi^2,$$

with boundary values

$$\xi^0(-1) = 1, \quad \xi^{0,\pm}(1) = 3, \quad \lim_{h\to\infty} \xi^{\pm}(h) = \infty.$$

iii. For $h \in (-1,1)$, ξ^0 is strictly monotonic increasing.

iv. For $h \in (1,\infty)$, $3h < \xi^{\pm}(h) \leqq 4h$ and $\xi^{\pm}$ is strictly monotonic increasing.

In each numbered section below we prove the corresponding property of ξ.

3i. For $h \in (-1,1)$ observe that Γ_h^0 is contractible to the point $(0,0) \in TS^1$. Therefore

$$\mathscr{A}^0 = \int_{\Gamma_h^0} d\phi = 0, \tag{3.1}$$

which implies that $\eta^0 = a\mathscr{A}^0/\mathscr{C}^0 - \xi^0 = -\xi^0$, provided that $\mathscr{C}^0$ is finite and nonzero. But

$$\mathscr{C}^0(h) = \int_{\Gamma_h^0} y\cos\phi\, d\phi = 2\sqrt{2}\int_{\phi_-}^{\phi_+} (h+\cos\phi)^{1/2}\cos\phi\, d\phi \tag{3.2}$$

(where $h + \cos\phi_{\pm} = 0$ and y is given in (1.3)), which implies that $\mathscr{C}^0(h)$ is finite for $h \in (-1,1)$. Differentiating (1.3) with respect to ϕ gives

$$\frac{dy}{d\phi} = -\frac{\sin\phi}{y} \tag{3.3}$$

and integrating (3.2) by parts yields

$$\mathscr{C}^0(h) = -\int_{\Gamma_h^0} \frac{dy}{d\phi}\sin\phi\, d\phi = \int_{\Gamma_h^0} \frac{\sin^2\phi}{y}\, d\phi = \sqrt{2}\int_{\phi_-}^{\phi_+} \frac{\sin^2\phi}{(h+\cos\phi)^{1/2}}\, d\phi. \tag{3.4}$$

From (3.4) it follows that $\mathscr{C}^0(h)>0$ for $h\in(-1,1)$. This proves that $\eta=-\xi^0$, and also we have shown that ξ^0 is continuous. To show that $\xi^-=\xi^+$, we demonstrate that

$$\mathscr{B}^+=\mathscr{B}^-, \qquad \mathscr{C}^+=\mathscr{C}^-. \tag{3.5}$$

This follows from the fact that y is positive on Γ_h^+, while it is negative on Γ_h^-. Because Γ^+ and Γ^- have opposite orientations, $\mathscr{B}$ and $\mathscr{C}$ are invariant under the $\pm$-change. For future reference, we note that the last argument also shows that

$$\mathscr{A}^+=-\mathscr{A}^-. \tag{3.6}$$

3ii. We begin by deriving the Riccati equation satisfied by $\xi=\xi^{0,\pm}$. In what follows all integrals are taken over the closed cycle $\Gamma_h^{0,\pm}$. Differentiating $\mathscr{B}=\mathscr{B}^{0,\pm}$ and $\mathscr{C}=\mathscr{C}^{0,\pm}$ with respect to h gives

$$\frac{d\mathscr{B}}{dh}=\int\frac{dy}{dh}d\phi=\int\frac{1}{y}d\phi \tag{3.7}$$

and

$$\frac{d\mathscr{C}}{dh}=\int\frac{dy}{dh}\cos\phi\,d\phi=\int\frac{1}{y}\cos\phi\,d\phi \tag{3.8}$$

(since differentiating (1.3) with respect to h yields $dy/dh=1/y$). Integrating $\mathscr{C}$ by parts and using (3.3) gives

$$\begin{aligned}\mathscr{C}&=\int y\cos\phi\,d\phi=-\int\frac{dy}{d\phi}\sin\phi\,d\phi=\int\frac{\sin^2\phi}{y}d\phi\\&=\int\frac{1}{y}\left[1-\left(\frac{1}{2}y^2-h\right)^2\right]d\phi\\&=\int\left[\frac{1}{y}-\frac{1}{4}y^3+hy-\frac{h^2}{y}\right]d\phi\\&=\int\left[\frac{1}{y}-\frac{1}{2}y(h+\cos\phi)+hy-\frac{h^2}{y}\right]d\phi\\&=(1-h^2)\frac{d\mathscr{B}}{dh}+\frac{1}{2}h\mathscr{B}-\frac{1}{2}\mathscr{C},\end{aligned} \tag{3.9}$$

which implies

$$(1-h^2)\frac{d\mathscr{B}}{dh}=-\frac{1}{2}h\mathscr{B}+\frac{3}{2}\mathscr{C}. \tag{3.10}$$

Similarly

$$\frac{d\mathscr{C}}{dh}=\int\frac{\cos\phi}{y}d\phi=\int\frac{1}{y}\left(\frac{1}{2}y^2-h\right)d\phi=\frac{1}{2}\mathscr{B}-h\frac{d\mathscr{B}}{dh},$$

which implies

$$(1-h^2)\frac{d\mathscr{C}}{dh}=\frac{1}{2}\mathscr{B}-\frac{3}{2}h\mathscr{C}. \tag{3.11}$$

Therefore the Picard–Fuchs equation for $\mathscr{B}$ and $\mathscr{C}$ is

$$2(1-h^2)\frac{d}{dh}\begin{pmatrix}\mathscr{B}\\ \mathscr{C}\end{pmatrix}=\begin{pmatrix}-h & 3\\ 1 & -3h\end{pmatrix}\begin{pmatrix}\mathscr{B}\\ \mathscr{C}\end{pmatrix}. \tag{3.12}$$

Since $\xi=\mathscr{B}/\mathscr{C}$, the Riccati equation satisfied by $\xi=\xi^{0,\pm}$ is

$$2(1-h^2)\frac{d\xi}{dh}=2(1-h^2)\frac{1}{\mathscr{C}}\frac{d\mathscr{B}}{dh}-\frac{\mathscr{B}}{\mathscr{C}}2(1-h^2)\frac{1}{\mathscr{C}}\frac{d\mathscr{B}}{dh}=3+2h\xi-\xi^2. \tag{3.13}$$

Next we compute the values of $\xi^{0,\pm}$ at $h=\pm 1$. We begin with the value of ξ^0 at $h=1$. As $h\uparrow 1$, the smooth closed path Γ_h^0 converges to the nonsmooth, but closed, path

$$\Gamma_1^0:\frac{1}{2}y^2=1+\cos\phi=2\cos^2\frac{\phi}{2},\qquad \phi\in[-\pi,\pi].$$

Therefore $\mathscr{B}^0$ converges to

$$\mathscr{B}^0(1)=\int_{\Gamma_1^0}y\,d\phi=4\int_{-\pi}^{\pi}\left|\cos\frac{\phi}{2}\right|d\phi=16, \tag{3.14}$$

while $\mathscr{C}^0$ converges to

$$\mathscr{C}^0(1)=\int_{\Gamma_1^0}\frac{\sin^2\phi}{y}d\phi=4\int_{-\pi}^{\pi}\sin^2\frac{\phi}{2}\left|\cos\frac{\phi}{2}\right|d\phi=\frac{16}{3}. \tag{3.15}$$

Therefore

$$\xi^0(1)=3. \tag{3.16}$$

As $h\downarrow -1$, Γ_h^0 shrinks to the point $(0,0)$. Let D_h^0 be the disk bounded by Γ_h^0. Then $\Gamma_h^0=\partial D_h^0$. Hence by Stokes' theorem

$$\xi^0(-1)=\lim_{h\to -1}\frac{\int_{\Gamma_h^0}y\,d\phi}{\int_{\Gamma_h^0}y\cos\phi\,d\phi}=\frac{\lim_{h\to -1}\int_{D_h^0}dy\,d\phi}{\lim_{h\to -1}\int_{D_h^0}\cos\phi\,dy\,d\phi}=1. \tag{3.17}$$

To compute $\xi^{\pm}$ at $h=1$, we note that the limit of $\Gamma_h^{\pm}$ at $h=1$ is given by

$$\Gamma_1^+:\frac{1}{2}y^2=1+\cos\phi=2\cos^2\frac{\phi}{2},\qquad \phi\in[-\pi,\pi],\quad y\geqq 0,$$

$$\Gamma_1^-:\frac{1}{2}y^2=1+\cos\phi=2\cos^2\frac{\phi}{2},\qquad \phi\in[-\pi,\pi],\quad y\leqq 0$$

(we may write $\Gamma_1^0=\Gamma_1^+\cup\Gamma_1^-$). Thus

$$\mathscr{B}^{\pm}(1)=8,\qquad \mathscr{C}^{\pm}(1)=\frac{8}{3} \tag{3.18}$$

and consequently

$$\xi^{\pm}(1)=3. \tag{3.19}$$

Then ξ is continuous at $h=1$, if we extend ξ^0 with either ξ^+ or ξ^-. Finally to show that $\lim_{h\to\infty}\xi^+(h)=\infty$, we need estimates for $\mathscr{B}^+$ and $\mathscr{C}^+$. When $h>1$ we obtain

$$\mathscr{C}^+(h)=\int_{\Gamma_h^+}\frac{\sin^2\phi}{y}d\phi=\frac{1}{\sqrt{2}}\int_{-\pi}^{\pi}\frac{\sin^2\phi}{(h+\cos\phi)^{1/2}}d\phi<\sqrt{2}\,\pi\frac{1}{\sqrt{h-1}} \tag{3.20}$$

and

$$\mathscr{B}^+(h)=\int_{\Gamma_h^+} y\,d\phi=\sqrt{2}\int_{-\pi}^{\pi}(h+\cos\phi)^{1/2}d\phi>2\pi\sqrt{2}\sqrt{h-1}\,. \tag{3.21}$$

Therefore when $h>1$

$$\xi^+(h)=\frac{\mathscr{B}^+(h)}{\mathscr{C}^+(h)}>\frac{2\pi\sqrt{2}\sqrt{h-1}}{\pi\sqrt{2}}\sqrt{h-1}=2(h-1) \tag{3.22}$$

which implies that $\lim_{h\to\infty}\xi^+(h)=\infty$.

3iii. To show that $\xi=\xi^0$ is strictly monotonic increasing we give an argument analogous to that used in Cushman and Sanders [9]. First observe that the connected components of

$$3+2h\xi-\xi^2=0 \tag{3.23}$$

are strictly monotonic functions of h, because differentiating (3.23) with respect to h and putting $d\xi/dh=0$ gives $\xi=0$. But this is ruled out by (3.23).

We now use this observation to show that the range of ξ^0 on $[-1,1]$ is $[1,3]$. From (3.13) it follows that on $(-1,1)$

$$\text{when } 0<\xi<1 \text{ then } \frac{d\xi}{dh}>0 \tag{3.24a}$$

while

$$\text{when } \xi>3 \text{ then } \frac{d\xi}{dh}<0. \tag{3.24b}$$

We now argue by contradiction. Suppose that for some $h_0\in(-1,1)$ we have $\xi(h_0)>3$. Since $\xi(-1)=1$, $\xi(1)=3$ and ξ is continuous, there is an $h_1\in(-1,h_0)$ such that $\xi(h_1)=3$. Since (3.13) implies that ξ is continuously differentiable on $(-1,1)$, there is an $h_2\in(h_1,h_0)$ such that $\xi(h_2)>0$ and $(d\xi/dh)(h_2)>0$. But this contradicts (3.24b). Therefore $\xi(h)\leqq 3$ for $h\in(-1,1)$. By an analogous argument one proves that $\xi(h)\geqq 1$ for $h\in(-1,1)$. Thus we have $1\leqq\xi(h)\leqq 3$ for all $h\in[-1,1]$.

Next we show that the derivative of ξ on $(-1,1)$ does not vanish. Suppose the contrary, that is, suppose that for some $h_0\in(-1,1)$, $(d\xi/dh)(h_0)=0$. Then differentiating (3.13) with respect to h and evaluating at h_0 gives

$$(1-h_0^2)\frac{d^2\xi}{dh^2}(h_0)=\xi(h_0). \tag{3.25}$$

Since the range of ξ on $[-1,1]$ is $[1,3]$, $(d^2\xi/dh^2)(h_0)>0$. In other words, we have shown that any extremum of ξ on $(-1,1)$ is a nondegenerate minimum. Because $\xi(-1)=1<\xi(h_0)$, ξ must have a maximum between -1 and h_0. But every extremum is a minimum on $(-1,1)$; so we have a contradiction. Consequently the derivative of ξ does not vanish on $(-1,1)$. Since $\xi(-1)=1$ and $\xi(1)=3$, $d\xi/dh>0$ on $(-1,1)$, that is, ξ is strictly monotonic increasing on $(-1,1)$.

3iv. Applying the change of variables

$$x=\frac{1}{\xi^+},\qquad t=\frac{1}{h} \tag{3.26}$$

to (3.13) gives

$$2t(1-t^2)\frac{dx}{dt} = t - 2x - 3tx^2. \tag{3.27}$$

We know that $\xi^+(1)=3$ and $\lim_{h\to\infty}\xi^+(h)=\infty$, so

$$x(0)=0 \quad \text{and} \quad x(1)=\frac{1}{3}.$$

First we prove that $x(t)\leqq t/3$ for $t\in[0,1]$. Let $y=x-t/3$, then $y(0)=y(1)=0$. From (3.27) it follows that

$$2t(1-t^2)\frac{dy}{dt} = -3ty^2 - 2y(1+t^2) - \frac{t}{3}(1-t^2). \tag{3.28}$$

Suppose that $y(\tau)>0$ for some $\tau\in(0,1)$. Then there is a $\tau_0\in(0,\tau)$ with $(dy/d\tau)(\tau_0)>0$ and $y(\tau_0)>0$; but this is in contradiction with the differential equation (3.28). Thus $y(t)\leqq 0$ for all $t\in[0,1]$, that is,

$$x(t)\leqq \frac{t}{3}. \tag{3.29}$$

Consequently by (3.24)

$$\xi^+(h)\geqq 3h \quad \text{for all } h\in[1,\infty). \tag{3.30}$$

Now suppose that $y(\tau)=0$ for some $\tau\in(0,1)$. Then from (3.28) we conclude that $(dy/dt)(\tau)<0$. But this contradicts the fact that $y(t)\leqq 0$ for all $t\in[0,1]$. Therefore

$$\xi^+(h)>3h \quad \text{for all } h\in(1,\infty). \tag{3.31}$$

The following argument shows that ξ^+ is strictly monotonic increasing on $(1,\infty)$. Since the zero isocline $3+2h-\xi^2=0$ of the Riccati equation (3.13) is strictly monotonic increasing for $h\in(1,\infty)$, ξ^+ can intersect this isocline at most once. From (3.25) it follows that ξ^+ has a nondegenerate (global) maximum at this intersection. This contradicts (3.31). Therefore $d\xi^+/dh>0$ on $(1,\infty)$. Consequently ξ^+ is strictly monotonic increasing on $(1,\infty)$.

Next we show that $\xi^+(h)\leqq 4h$ on $[1,\infty)$. Toward this end, let $y=x-t/4$; then $y(0)=0$, $y(1)=\frac{1}{12}$ and (3.27) becomes

$$2t(1-t^2)\frac{dy}{dt} = -3ty^2 - \left(2+\frac{3}{2}t^2\right)y + \frac{5}{16}t^3. \tag{3.32}$$

Suppose that for some $\tau\in(0,1)$, $y(\tau)<0$. Then there is a $\tau_0<(0,\tau)$ with $(dy/dt)(\tau_0)<0$ and $y(\tau_0)<0$. But, by (3.29)

$$y(t)=x(t)-\frac{t}{4}\leqq \frac{t}{3}-\frac{t}{4}=\frac{t}{12},$$

which implies

$$2t(1-t^2)\frac{dy}{dt} \geqq \frac{7}{24}t^3 - \left(2+\frac{3}{2}t^2\right)y. \tag{3.33}$$

Therefore $(dy/dt)(\tau_0)>0$, which is a contradiction. Thus $y(t)\geqq 0$ for all $t\in[0,1]$ or equivalently

$$x(t)\geqq \frac{t}{4}, \tag{3.34}$$

that is,

$$\xi^+(h) \leqq 4h \quad \text{for all } h \in [1, \infty). \tag{3.35}$$

Combining (3.31) and (3.35), we find that

$$3h < \xi^+(h) \leqq 4h \quad \text{for all } h \in (1, \infty). \tag{3.36}$$

This completes the proof of the properties of the function ξ.

4. Properties of the function η. Throughout this section we assume that $h > 1$. We will show that the function $\eta = \eta^{\pm}$, where

$$\eta = a\frac{\mathscr{A}}{\mathscr{C}} - \xi \tag{4.1}$$

has the following properties:

(4.2)
i. $\eta^{\pm}(1) = \pm 3\pi a/4 - 3$.
ii. $\eta = \eta^{\pm}$ satisfies the Riccati equation

$$2(h^2-1)\frac{d\eta}{dh} = 3 - h\xi - 3h\eta + \xi\eta \tag{4.3}$$

where $\xi = \xi^{\pm}$.

iii. Let

$$\zeta(h) = \frac{h\xi - 3}{\xi - 3h}; \tag{4.4}$$

then $\zeta(1) = 1$. In addition, the zero isocline $\eta^{\pm} = \xi$ of (4.3), is continuous and strictly monotonic increasing.

iv. If $\eta^{\pm}(1) > 1$, that is, if

$$a > \frac{16}{3\pi}, \quad \eta = \eta^+ \quad \text{or} \quad a < -\frac{16}{3\pi}, \quad \eta = \eta^-,$$

then $\eta^{\pm}$ has a unique maximum $\tilde{\eta}^{\pm}(a)$ when $h > 1$; moreover $\lim_{a \to \pm 16/3\pi} \tilde{\eta}^{\pm}(a) = 1$. If $\eta^{\pm}(1) < 1$, then $\eta^{\pm}$ is strictly monotonic decreasing on $(1, \infty)$.

As before, in each numbered section we prove the corresponding property of η.

4i. Since $\mathscr{C}^+ = \mathscr{C}^-$, $\xi^+ = \xi^-$ and $\mathscr{A}^+ = -\mathscr{A}^-$ with $\mathscr{C}^+(1) = \frac{8}{3}$ (3.18), $\xi^+(1) = 3$ (3.19), and $\mathscr{A}^+(h) = 2\pi$ we get

$$\eta^+(1) = \frac{3\pi}{4}a - 3 \quad \text{and} \quad \eta^-(1) = -\frac{3\pi}{4}a - 3. \tag{4.5}$$

4ii. Let $\eta = \eta^{\pm}$. Then differentiating (4.1) with respect to h using (3.11), (3.13) and some regrouping gives

$$2(h^2-1)\frac{d\eta}{dh} = 3 - h\xi - 3h\eta + \xi\eta. \tag{4.6}$$

Here the fact that $\mathscr{A}^{\pm}$ is constant on $(1, \infty)$ has been used.

4iii. The zero isocline of (4.3) is the locus of zeros of the right-hand side of (4.6) which is given by $\eta = \eta^{\pm} = \zeta$ where

$$\zeta(h) = \frac{h\xi - 3}{\xi - 3h}. \tag{4.7}$$

Since $\xi > 3h$ when $h > 1$ (3.31), we see that ζ is continuous on $(1, \infty)$. To show that ζ is strictly monotonic increasing on $(1, \infty)$, we differentiate (4.7) with respect to h and obtain

$$\frac{d\zeta}{dh} = -\frac{1}{2(\xi - 3h)^2}(9 - 6h\xi + \xi^2). \tag{4.8}$$

Put $\theta(h) = 9 - 6h\xi + \xi^2$. Since $\xi(1) = 3$, $\theta(1) = 0$; moreover, since $3h < \xi(h) \leqq 4h$ for $h \in [1, \infty)$,

$$\theta(h) \leqq 9 - 2h^2 \quad \text{on } [1, \infty). \tag{4.9}$$

Therefore $\theta(h) < 0$ for $h > \frac{3}{2}\sqrt{2}$. To show that $d\zeta/dh > 0$ it suffices to show that $\theta < 0$ for all $h > 1$. Toward this end we differentiate θ with respect to h. Repeatedly using $\xi^2 = \theta + 6h\xi - 9$ we obtain

$$(h^2 - 1)\frac{d\theta}{dh} = (h + \xi)\theta + 6(h^2 - 1)\xi, \tag{4.10}$$

which may be rewritten as

$$\frac{d\theta}{dh} = 6\xi + \frac{h + \xi}{h^2 - 1}\theta. \tag{4.11}$$

Suppose that there is an $h_0 > 1$ such that $\theta(h_0) \geqq 0$. Since $\theta(h) < 0$ for all $h > \frac{3}{2}\sqrt{2}$, there is an $h_1 \geqq h_0$ such that $\theta(h_1) = 0$. Take h_1 to be the largest value of h such that $\theta(h_1) = 0$. Then $\theta(h) < 0$ for all $h > h_1$. It follows from (4.11) and (3.31) that

$$\frac{d\theta}{dh}(h_1) \geqq 6\xi(h_1) > 18h_1. \tag{4.12}$$

Thus there is an $h_2 > h_1$, for which $\theta(h_2) = 0$; but this contradicts the maximality of h_1. Therefore $\theta(h) < 0$ for $1 < h < \infty$, which implies that ζ is strictly monotonic increasing on $(1, \infty)$.

To compute $\lim_{h \to 1^+} \zeta(h)$, we first have to show that

$$\lim_{h \to 1^+} \xi'(h) = \infty. \tag{4.13}$$

The following argument shows that either $\lim \xi' < \infty$ or $\lim \xi' = \infty$ as $h \to 1^+$. Suppose not, then

$$\liminf \xi' < \limsup \xi' \tag{4.14a}$$

and

$$-\infty < \liminf \xi' < \infty. \tag{4.14b}$$

Since ξ is monotonic increasing on $(1, \infty)$, we may replace (4.14b) with

$$0 \leqq \liminf \xi' < \infty. \tag{4.15}$$

From (4.14a) there is a sequence $\{h_n\}$, $h_n \to 1^+$ such that ξ' has a local minimum at h_n, that is, $\xi''(h_n) = 0$. Differentiating the Riccati equation (3.13) gives

$$2(1 - h^2)\xi'' = 2(3h - \xi)\xi' + 2\xi \tag{4.16a}$$

which evaluated at h_n gives

$$0=2(3h_n-\xi(h_n))\xi'(h_n)+2\xi(h_n). \tag{4.16b}$$

Taking the lim inf as $h_n\to 1^+$ of both sides of (4.16a) gives

$$0=\liminf_{h_n\to 1^+}\xi'\lim_{h_n\to 1^+}(3h_n-\xi(h_n))+\lim_{h_n\to 1^+}\xi(h_n) \tag{4.17}$$

since ξ is continuous, the first term in (4.17) vanishes because $\xi(1)=3$ and (4.15), while the second term equals 3. Thus (4.17) is false. Consequently either $\lim_{h\to 1^+}\xi'<\infty$ or $\lim_{h\to 1^+}\xi'=\infty$. Now suppose that $\lim_{h\to 1^+}\xi'<\infty$. Then the Riccati equation (3.13) gives the asymptotic relation

$$\xi''\sim\frac{1}{(1+h)(1-h)}\xi\sim\frac{3}{2(1-h)}\quad\text{as } h\to 1^+$$

which integrated shows that $\xi'\to\infty$ as $h\to 1^+$. This contradicts the hypothesis. Therefore (4.13) holds.

Using l'Hôpital's rule and (4.12), we compute $\zeta(1)$ as follows:

$$\zeta(1)=\lim_{h\to 1^+}\frac{h\xi-3}{\xi-3h}=\lim_{h\to 1^+}\frac{\xi(h)+h\xi'(h)}{\xi'(h)-3}=1. \tag{4.18}$$

4iv. Let $\eta=\eta^+$. From (4.3) it follows that $d\eta/dh>0$ above the graph of ζ, while $d\eta/dh<0$ below. Thus when $\eta(1)<1$, η is strictly monotonic decreasing for $h>1$. In addition the range of η is $(-\infty,\eta(1))$, as we shall prove below (see (4.21)).

When $\eta(1)>1$ and $h>1$, η is strictly monotonic increasing as long as its graph does not intersect the graph of ζ. But this can happen at most once, because η is strictly monotonic decreasing after such an intersection and ζ is strictly monotonic increasing when $h>1$.

To show that the graphs of η and ζ intersect when $h>1$, we show that $\lim_{h\to\infty}\eta(h)=-\infty$, whatever the initial condition $\eta(1)$. First we find the asymptotic behavior of $\mathscr{C}^+$:

$$\begin{aligned}\mathscr{C}^+(h)&=\int_{\Gamma_h^+}\frac{\sin^2\phi}{y}d\phi=\frac{1}{\sqrt{2h}}\int_{-\pi}^{\pi}\sin^2\phi\left(1+h^{-1}\cos\phi\right)^{-1/2}d\phi\\&=\frac{1}{2h}\left[\int_{-\pi}^{\pi}\left(\sin^2\phi+O\left(\frac{1}{h}\right)\right)d\phi\right]=\frac{\pi}{\sqrt{2h}}+O\left(\frac{1}{h^{3/2}}\right).\end{aligned} \tag{4.19}$$

From the definition of $\eta=\eta^{\pm}$ (1.7) we obtain

$$\eta^{\pm}(h)=\frac{a\mathscr{A}^{\pm}}{\mathscr{C}^{\pm}}-\xi^{\pm}=\frac{2\pi a}{\mathscr{C}^{\pm}(h)}-\xi^{\pm}(h)=\pm\frac{2\pi a}{\left(\pi/\sqrt{2h}+O(1/h^{3/2})\right)}-\xi^{\pm}(h). \tag{4.20}$$

Since $3h<\xi(h)\leqq 4h$ it follows that

$$\lim_{h\to\infty}\eta^{\pm}(h)=-\infty. \tag{4.21}$$

This implies that the range of η is $(-\infty,\eta(1))$ if $\eta(1)<1$ and $(-\infty,\tilde{\eta}(a))$ if $\eta(1)>1$, where $\tilde{\eta}(a)$ is the maximum value of η.

Because the range of $\eta^{\pm}$ is $(-\infty, 1]$ when $a = \pm 16/3\pi$, by continuity it follows that the maximum value of η^{+}, respectively η^{-}, for $a > 16/3\pi$, respectively, $a < -16/3\pi$, lies in the range $(1, \infty)$. Thus

$$\lim_{a \to (16/3\pi)^{+}} \tilde{\eta}^{+}(a) = \lim_{a \to -(16/3\pi)^{-}} \tilde{\eta}^{-}(a) = 1.$$

This completes the proof of the properties of η. Thus we have proved the main theorem stated in §1.

5. Appendix: Asymptotic expansions for ξ and η when h is slightly greater than 1. The maximum value of η as a function of a has to be computed numerically. This proved difficult to do for a near $16\pi/3$. In this appendix we give asymptotic approximations for η to explain this difficulty and to provide an alternative proof that

$$\lim_{a \to (16\pi/3)^{+}} \tilde{\eta}^{+}(a) = 1.$$

Consider the Riccati equation (3ii)

$$2(1-h^2)\frac{d\xi}{dh} = 3 + 2h\xi - \xi^2. \tag{5.1}$$

Since $\xi(1) = 3$, taking the limit as $h \to 1^{+}$ of both sides of (5.1) gives

$$\lim_{h \to 1^{+}} (1-h)\frac{d\xi}{dh} = 0. \tag{5.2}$$

Now we deduce an asymptotic relation for ξ as $h \to 1^{+}$. Differentiating (5.1) with respect to h gives

$$(1-h)(1+h)\frac{d^2\xi}{dh^2} = \xi + (3h - \xi)\frac{d\xi}{dh}. \tag{5.3}$$

Because $\xi(1) = 3$ and (5.2), from (5.3) we obtain the asymptotic relation

$$\frac{d^2\xi}{dh^2} \sim \frac{-3}{2(h-1)} \quad \text{as } h \to 1^{+}. \tag{5.4}$$

Integrating (5.4) twice gives

$$\xi(h) \sim 3 - \frac{3}{2}(h-1)\log(h-1), \qquad h \to 1^{+} \tag{5.5}$$

as desired.

Next we deduce an asymptotic relation for $\eta \to 1^{+}$. Substituting (5.5) into the Riccati equation (4.3),

$$2(h^2-1)\frac{d\eta}{dh} = 3 - h\xi - 3h\eta + \xi\eta, \tag{5.6}$$

gives

$$2(h-1)(h+1)\frac{d\eta}{dh} \sim -3(h-1) - 3\eta(1)(h-1) + \frac{3}{2}(h - \eta(1))(h-1)\log(h-1).$$

that is,

$$\frac{d\eta}{dh} \sim -3(1+\eta(1)) + \frac{3}{2}(1-\eta(1))\log(h-1) \text{ as } h \to 1^{+}. \tag{5.7}$$

Integrating (5.7) gives the asymptotic relation

$$(5.8)\quad \eta(h)\sim\eta(1)-\frac{3}{8}(3+\eta(1))(h-1)+\frac{3}{8}(1-\eta(1))(h-1)\log(h-1)\quad \text{as } h\to 1^+$$

as desired. To compute an asymptotic relation for the critical point $h=h^*$ of η, we equate the right-hand side of (5.7) to zero. Thus

$$(5.9)\qquad h^*-1=\exp\left[\frac{2(1+\eta(1))}{1-\eta(1)}\right].$$

Now $\eta(1)=3\pi a/4-3$. Putting $a=(16/3\pi)(1+\varepsilon/4)$ gives $\eta(1)=1+\varepsilon$, which substituted into (5.9) yields

$$(5.10)\qquad h^*-1=e^{-2}e^{-4/\varepsilon}.$$

Substituting (5.10) into (5.8) gives the asymptotic relation

$$(5.11)\qquad \eta_{\max}=\eta(h^*)\sim\eta(1)-\frac{3}{8}(4+\varepsilon)e^{-2}e^{-4/\varepsilon}+\frac{3}{8}\left(\frac{4}{\varepsilon}+2\right)e^{-2}e^{-4/\varepsilon}$$

$$=\eta(1)+\frac{3}{8}\varepsilon e^{-2}e^{-4/\varepsilon}\quad \text{as } \varepsilon\to 0^+.$$

Therefore the difference between the maximum value of η for a slightly greater than $16/3\pi$ and the boundary value $\eta(1)$ is exponentially small. This explains the difficulty in computing the maximum of η near the critical value $a=16/3\pi$.

Acknowledgments. We would like to thank Prof. M. van Veldhuizen of the Free University in Amsterdam for doing the numerical calculations which resulted in Fig. 2. We would also like to thank the referee for the proof of (4.13) and helpful comments which greatly improved the organization and exposition of this paper.

REFERENCES

[1] V. I. ARNOL'D, *Lectures on bifurcations in versal families*, Russian Math. Surveys, 27 (1972), pp. 54–123.

[2] _____, *Geometrical Methods in the Theory of Ordinary Differential Equations*, Springer-Verlag, New York, 1983.

[3] V. N. BELYKH, N. F. PEDERSON AND O. H. SOERENSON, *Shunted-Josephson-junction model* I. *The autonomous case*, Phys. Rev. B, 16 (11) (1977), pp. 4853–4859.

[4] R. I. BOGDANOV, *Orbital equivalence of singular points of vector fields*, Funct. Anal. Appl., 10 (1976), pp. 316–317.

[5] _____, *Bifurcation of the limit cycle of a family of plane vector fields*, Selecta Math. Sov., 1 (1981), pp. 373–387.

[6] _____, *Versal deformation of a singularity of a vector field on the plane in the case of zero eigenvalues*, Selecta Math. Sov., 1 (1981), pp. 389–421.

[7] J. CARR, *Applications of Centre Manifold Theory*, Applied Mathematical Sciences 35, Springer-Verlag, New York, 1981.

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

[9] R. CUSHMAN AND J. A. SANDERS, *A codimension two bifurcation with a third order Picard–Fuchs equation*, J. Differential Equations, 59 (1985), pp. 243–256.

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

[11] E. I. HOROZOV, *Versal deformations of equivariant vector fields with* $\mathbb{Z}_2$ *or* $\mathbb{Z}_3$ *symmetry*, Trudy Seminara im. I. G. Petrovskovo, 5 (1979), pp. 163–192. (In Russian.)

[12] YU. S. IL'YASHENKO, *The multiplicity of limit cycles arising from perturbations of the form* $w'=P_2/Q_1$ *of a Hamiltonian equation in the real and complex domain*, Amer. Math. Soc. Transl., (2) 118 (1982), pp. 191–202.

[13] _____, *Zeros of special abelian integrals in a real domain*, Funct. Anal. Appl., 11 (1977), pp. 309–311.
[14] R. L. KAUTZ, *On a proposed Josephson effect voltage standard at zero current bias*, Appl. Phys. Lett., 36 (1980), pp. 386–388.
[15] L. S. PONTRJAGIN, *Über Autoschwingungssysteme, die den Hamiltonsche nahe liegen*, Phys. Z. Sowjetunion, 6 (1934), pp. 25–28.
[16] J. A. SANDERS, *Melnikov's method and averaging*, Celestial Mechanics, 28 (1982), pp. 171–181.
[17] _____, *The (driven) Josephson equation: an exercise in asymptotics*, in Asymptotic Analysis II, F. Verhulst, ed., Lecture Notes in Mathematics 985, Springer-Verlag, Berlin, 1983.
[18] G. S. PETROV, *On the number of zeros of the full elliptic integrals*, Funct. Anal. Appl., 18 (1982), pp. 148–149.
[19] R. J. SOULEN AND R. P. GIFFARD, *Impedance and noise of an rf-biased* RSQUID *operated in a nonhysteretic regime*, Appl. Phys. Lett., 32 (1978), p. 770.

SIAM J. MATH. ANAL.
Vol. 17, No. 3, May 1986

BREAKDOWN OF STABILITY IN SINGULARLY PERTURBED AUTONOMOUS SYSTEMS I. ORBIT EQUATIONS*

K. NIPP†

Abstract. In singular perturbation analysis two stages can generally be distinguished: First, by some method a sequence of formal approximations is constructed which are supposed to form together an asymptotic approximation to the solution of the singular perturbation problem. Second, it should be proved that the sequence provides a correct approximation in a rigorous mathematical sense.

For this second stage analytical results providing error estimates are needed. A classical one is due to A. N. Tikhonov (see e.g. [10]). We quote this theorem, and state and prove results for the trajectories of a singularly perturbed autonomous system that are extensions of the Tikhonov Theorem in the neighborhood of a point where a certain stability assumption ceases to be valid.

1. Introduction.

Consider the autonomous system

$$\begin{aligned} \dot{x} &= f(x,y)+\varepsilon f^1(x,y,\varepsilon), \\ \varepsilon\dot{y} &= g(x,y)+\varepsilon g^1(x,y,\varepsilon), \end{aligned} \tag{1}$$

together with the initial conditions

$$x(0,\varepsilon)=x^0(\varepsilon), \qquad y(0,\varepsilon)=y^0(\varepsilon) \tag{2}$$

where x and y are m- and n-vectors, respectively, and ε is a small nonnegative parameter. We assume that f, f^1 and g, g^1 are sufficiently smooth with respect to all variables in the domain considered as are x^0, y^0 for $\varepsilon\in[0,\varepsilon_0]$, with $\varepsilon_0<1$.

The corresponding reduced problem is obtained by putting $\varepsilon=0$ and by dropping the initial data on y:

$$\begin{aligned} \dot{x} &= f(x,y), \qquad x(0)=x^0(0), \\ 0 &= g(x,y). \end{aligned} \tag{3}$$

We make the following assumption:

(4) There is a continuously differentiable vector function $\phi(x)$ defined in some domain $D\subset\mathbb{R}^m$ containing $x^0(\varepsilon)$ for $\varepsilon\in[0,\varepsilon_0]$ and a positive constant b such that $g(x,\phi(x))=0$ for $x\in D$ and all the eigenvalues of the Jacobian matrix $g_y(x,\phi(x))$ have real parts smaller than $-b$ for $x\in D$.

The set $(x,\phi(x))$ is called a reduced manifold of the system (1). The property (4) implies stability of this reduced manifold in the sense of Theorem 1 below.

Moreover, we suppose that the problem (3) has a continuously differentiable solution $(X(t), Y(t))$ for $t\in[0,t_1]$, lying in the stable reduced manifold (i.e. $\dot{X}(t)=f(X(t),\phi(X(t)))$, $X(0)=x^0(0)$; $Y(t)=\phi(X(t))$). Then the following theorem due to A. N. Tikhonov holds.

*Received by editors May 31, 1983, and in revised form August 6, 1984.

†Applied Mathematics, ETH-Zentrum, CH-8092 Zürich, Switzerland. This work was carried out while the author was at the Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, New York, 12181. This work was supported by a grant from the Schweizerischer Nationalfonds.

THEOREM 1. *If* $|y^0(\varepsilon)-Y(0)|$ *is sufficiently small, then for* ε *small enough the solution* $(x(t,\varepsilon), y(t,\varepsilon))$ *of* (1), (2) *exists and is unique for* $t\in[0,t_1]$ *and satisfies*

$$|x(t,\varepsilon)-X(t)|=O(\varepsilon) \quad \textit{uniformly for } t\in[0,t_1]$$

and

$$|y(t,\varepsilon)-Y(t)|=O(\varepsilon) \quad \textit{uniformly for } t\in[\tau,t_1],\ \textit{any } \tau\in(0,t_1).$$

The above formulation is analogous to the one given in [2] or [5] with sharper estimates than in [10]. Of course, a similar result holds in the nonautonomous case.

In this paper we are interested in situations where the stability property (4) ceases to be valid. One of the implications of this condition is that the reduced manifold $y=\phi(x)$ is locally unique. This excludes bifurcation of the reduced manifold. In [3] and [4] Lebovitz and Schaar considered the two kinds of bifurcations sketched in Fig. 1, with the components of the reduced manifold being locally linear or quadratic, respectively, in the essential variables and having the indicated stability properties. At the origin, which is the point of bifurcation, exactly one single eigenvalue of the Jacobian is supposed to vanish.

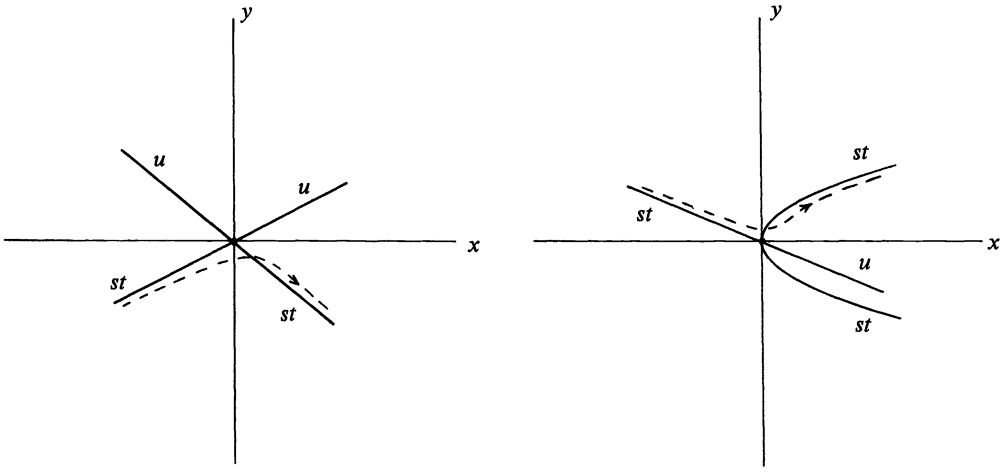


FIG. 1

Another bifurcation case is the so-called jump point situation sketched in Fig. 2, arising frequently in the context of relaxation oscillations. In both figures, the dashed lines represent a possible trajectory of the system (1). In the case of Fig. 2, typically, the trajectory follows the stable branch of the reduced manifold until it reaches a vicinity of the bifurcation point where it drops off.

There are several papers in the Russian literature dealing with this situation (see e.g. [8], [9]). It is also treated in the book by Mishchenko and Rozov [5] in the context of relaxation oscillations. These authors, however, consider only reduced manifolds which are locally quadratic in the essential variables. Moreover, they are primarily interested in working out the forms of the approximations to the IVP (1), (2) to the highest possible order rather than in giving a rigorous proof of their validity. There are error estimates provided in [9], [5] for the higher dimensional case which do seem not to be optimal, however.

In this paper our goal is to obtain sharp estimates for the domain of validity as well as for the error of an approximation given by the solution of the reduced system (3) in the neighborhood of a point where the stability property (4) is weakened. Therefore, the results stated below are extensions of Theorem 1.

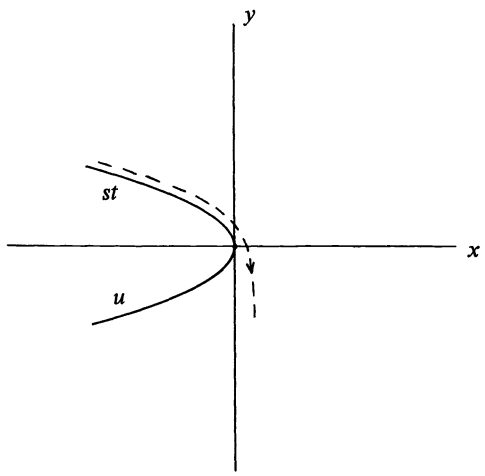


FIG. 2

It may be worthwhile to discuss some conceptual points. Rather than treat some specific problem we propose to derive certain general local results that permit to establish local estimates which, if taken together with Tikhonov's Theorem as well as some trivial estimates from regular perturbation theory, allow to provide global estimates for various bifurcation situations. In particular, the two kinds of bifurcations sketched in Figs. 1 and 2 turn out to be special cases with the reduced manifold being locally linear or quadratic.

We will first give a precise formulation of the problem. There are two cases which are treated separately. In the first one the reduced trajectory encounters a point where stability breaks down (as e.g. in Fig. 1 or Fig. 2 for $x<0$). The second case is characterized by the fact that the reduced trajectory departs from this point, a situation which arises in bifurcations of the type sketched in Fig. 1 ($x>0$). In this first paper we state and prove the results for the *orbits*. The transfer of the estimates to the *solutions* is postponed to Part II also to appear in this journal. There we will as well give an application of our results and of the concept mentioned above.

A two-dimensional version of the first of our results was given in [7]. As far as the construction of suitable local approximations is concerned the reader is referred to [6].

2. Formulation of the problem for the case "$s<0$". In this section we give a detailed formulation of the problem for the first case mentioned at the end of the previous section. This is in a sense the easier and more natural situation since the initial conditions for the local problem are provided by Theorem 1. We take the point, where the stability assumption (4) ceases to hold, to be the origin in $\mathbb{R}^{m+n}$. And we suppose that $f(0,0)\neq 0$ so that the origin is not a critical point of the full system (1). Without loss of generality we may assume that

H1. $$f_1(0,0)>0,\ f_k(0,0)=0 \qquad (k=2,3,\cdots,m).$$

Hence, there exist a domain U containing the origin and a positive constant ρ such that $f_1(x,y)>\rho$ for $(x,y)\in U$. And we may introduce $s=x_1$ as the independent variable. By dividing the rest of the equations of (1) by the first one we obtain, for ε small enough, the following orbit equations

$$\begin{aligned}\frac{d\bar{x}}{ds}&=F(s,\bar{x},y)+\varepsilon F^1(s,\bar{x},y,\varepsilon),\\ \varepsilon\frac{dy}{ds}&=G(s,\bar{x},y)+\varepsilon G^1(s,\bar{x},y,\varepsilon),\end{aligned}\tag{5}$$

where $\bar{x}$ represents the last $m-1$ components of x; $F:=\bar{f}/f_1$, $G:=g/f_1$, F^1 and G^1 are defined for $(x,y)\in U$, and $F(0,0,0)=0$, $G(0,0,0)=0$. The initial conditions are

$$\bar{x}(s^0,\varepsilon)=\bar{x}^0(\varepsilon),\qquad y(s^0,\varepsilon)=y^0(\varepsilon)\tag{6}$$

where we suppose that $s^0(\varepsilon)=S^0+O(\varepsilon)$ with $S^0<0$ independent of ε.

We assume that the Jacobian matrix $g_y(0,0)$ has zero as a simple eigenvalue and all the other eigenvalues have negative real parts. Hence the same is true for G_y. By appropriate choice of basis in $\mathbb{R}^n$, we may suppose that

H2. $$G_y(0,0,0)=\begin{pmatrix}0 & 0 & \cdots & 0\\ 0 & & & \\ \vdots & & \bar{A} & \\ 0 & & & \end{pmatrix}$$

where all the eigenvalues of the $(n-1)\times(n-1)$ matrix $\bar{A}$ have negative real parts. Moreover, let $\bar{y}:=(y_2,\cdots,y_n)$. Then

$$\bar{G}(x,y_1,\bar{y})=\bar{A}\bar{y}+Bx+\bar{c}y_1^2+\hat{G}(x,y_1,\bar{y})\tag{7}$$

where the remainder term $\hat{G}$ is quadratic in (x,y) and cubic in y_1. Without loss of generality we may assume that $\bar{c}=0$. (This can always be achieved by a simple transformation of variables (cf. [4])).

H2 implies that there exists a unique solution $\bar{w}(s,\bar{x},y_1)$ of $\bar{G}(s,\bar{x},y_1,\bar{y})=0$ in a neighborhood of the origin, having continuous partial derivatives there and satisfying $\bar{w}(0,0,0)=0$. For the remaining equation

$$G_1(s,\bar{x},y_1,\bar{w}(s,\bar{x},y_1))=0\tag{8}$$

we require:

H3. There exists a function $\phi_1(s,\bar{x})$, which is defined and continuous for $s\in J:=[s^0,0]$, $\bar{x}$ in some neighborhood $\bar{\Omega}$ of $\bar{x}=0$, has continuous partial derivatives with respect to $\bar{x}$ there, and is C^1 in $[s^0,0)\times\bar{\Omega}$, satisfying

$$\begin{aligned}&G_1(s,\bar{x},\phi_1(s,\bar{x}),\bar{w}(s,\bar{x},\phi_1(s,\bar{x})))=0,\qquad (s,\bar{x})\in J\times\bar{\Omega},\\ &\phi_1(0,0)=0\end{aligned}\tag{9}$$

and

$$\phi_1(s,\bar{x})=a_1(-s)^{\alpha}+p(s,\bar{x}),\qquad (s,\bar{x})\in J\times\bar{\Omega},\tag{10}$$

$$\phi_{1,s}(s,\bar{x})=a_{11}(-s)^{\alpha-1}+p_1(s,\bar{x}),\qquad (s,\bar{x})\in[s^0,0)\times\bar{\Omega},\tag{11}$$

where $\alpha>0$ and a_1, $a_{11}\neq 0$, $p(0,0)=0$, $p(s,0)=o((-s)^{\alpha})$ and $p_1(s,0)=o((-s)^{\alpha-1})$ as $s\to 0^-$.

H3′. If $\alpha \geqq 1$ we suppose that $\phi_1(s,\bar{x}) \in C^1(J \times \bar{\Omega})$.

Remark 1. In the case $\alpha \geqq 1$ (H3′) (11) follows from (10), but not in the case $\alpha < 1$ since in general order relations may not be differentiated. An assumption implying (10), (11) and easy to verify would be: There exists a function $\hat{\phi}(\hat{s},\bar{x}) \in C^1(\Omega_1 \times \bar{\Omega})$ (Ω_1 an appropriate interval containing 0) with $\hat{\phi}(0,0)=0$, $\hat{\phi}_{\hat{s}}(0,0) \neq 0$ and $\phi_1(s,\bar{x}) = \hat{\phi}((-s)^{\alpha},\bar{x})$, $(s,\bar{x}) \in J \times \bar{\Omega}$. This is a natural condition for jump points.

Let now $\bar{\phi}(s,\bar{x}) := \bar{w}(s,\bar{x},\phi_1(s,\bar{x}))$ and $\phi(s,\bar{x}) := (\phi_1,\bar{\phi})$, and consider the initial value problem

$$\frac{d\bar{x}}{ds} = F(s,\bar{x},\phi(s,\bar{x})), \qquad \bar{x}(0)=0. \tag{12}$$

We suppose the following.

H4. The solution $\bar{X}(s)$ of (12) exists for $s \in J$ and there is $c_0 > 0$ such that

$$|\bar{x}^0(\varepsilon) - \bar{X}(s^0)| < c_0\varepsilon, \qquad |y^0(\varepsilon) - Y(s^0)| < c_0\varepsilon,$$

where $Y(s) := \phi(s,\bar{X}(s))$.

Remark 2. H4 is naturally satisfied (by means of Theorem 1) if we assume having a global problem (i.e. system (1) together with some initial conditions) which is of Tikhonov type away from the origin, or more precisely whose reduced trajectory is stable as long as it is in a finite distance from the origin and, moreover, is zero there. Thus, without loss of generality, we may assume that the initial conditions (6) are "small enough", i.e. such that they lie in the appropriate neighborhood of the origin considered here. In the following this will also be referred to by the expression "for $|s^0|$ small enough". In particular we assume $|s^0| < 1$.

We now state the essential stability condition which, together with H2, replaces assumption (4) for the Tikhonov case and which is much more general than the corresponding ones given in [3], [4], [5].

H5. There are positive constants k and q such that

$$G_{1,y_1}(s,\bar{X}(s),Y(s)) \leqq -k(-s)^q, \qquad s \in J.$$

Remark 3. In order to verify this assumption the asymptotic relations for $(\bar{X}(s), Y(s))$ given in the next section may be used. The estimates (27)–(29) might, however, be sharper in a given application. We will also derive relations between q and α stated in two lemmas in the next section.

3. Preliminaries. In this section we provide all estimates needed for proving the subsequent results.

Let the functions $u(s,\varepsilon)$ and $v(s,\varepsilon)$ be defined by means of

$$\bar{x}(s,\varepsilon) = \bar{X}(s) + u(s,\varepsilon), \qquad y(s,\varepsilon) = Y(s) + v(s,\varepsilon). \tag{13}$$

They satisfy the differential equations

$$\begin{aligned}
&\frac{du}{ds} = F_{\bar{x}}(s)u + F_y(s)v + R(s,u,v,\varepsilon),\\
&\varepsilon\frac{dv_1}{ds} = G_{1,\bar{x}}(s)u + G_{1,y_1}(s)v_1 + G_{1,\bar{y}}(s)\bar{v} + S_1(s,u,v,\varepsilon) - \varepsilon Y_1'(s),\\
&\varepsilon\frac{d\bar{v}}{ds} = \bar{G}_{\bar{x}}(s)u + \bar{G}_{y_1}(s)v_1 + \bar{G}_{\bar{y}}(s)\bar{v} + \bar{S}(s,u,v,\varepsilon) - \varepsilon\bar{Y}'(s),
\end{aligned} \tag{14}$$

where $F_{\bar{x}}(s) := F_{\bar{x}}(s,\bar{X}(s),Y(s))$ and similarly for the other coefficient functions.

For some choice of the positive constants $K>1$, $\theta_1 \geqq |s^0|$ and θ the remainder terms R, S_1 and $\bar{S}$ in (14) satisfy the following estimate

$$|R(s,u,v,\varepsilon)| \leqq K\left(\varepsilon + |u|^2 + |v|^2\right) \quad \left(\text{and similarly for } S_1 \text{ and } \bar{S}\right) \tag{15}$$

if $-\theta_1 \leqq s \leqq 0$, $|u|+|v|<\theta$, and $\varepsilon \in (0,\varepsilon_0)$. The common constant K is supposed to be so large that these and later estimates hold simultaneously. We will see below that it is possible to slightly improve the estimates of S_1 and $\bar{S}$.

As initial conditions to the system (14) we have

$$u(s^0,\varepsilon) = \bar{x}^0(\varepsilon) - \bar{X}(s^0), \qquad v(s^0,\varepsilon) = y^0(\varepsilon) - Y(s^0). \tag{16}$$

Let $U(s,\sigma)$, $V_1(s,\sigma,\varepsilon)$, $\bar{V}(s,\sigma,\varepsilon)$ be the fundamental matrix solutions of

$$\frac{dU}{ds} = F_{\bar{x}}(s)U, \quad \frac{dV_1}{ds} = \varepsilon^{-1}G_{1,y_1}(s)V_1, \quad \frac{d\bar{V}}{ds} = \varepsilon^{-1}\bar{G}_{\bar{y}}(s)\bar{V}, \tag{17}$$

reducing to the unit matrices of dimensions $m-1$, 1, $n-1$, respectively, when $s=\sigma$.

The initial value problem (14), (16) is equivalent to the following system of integral equations (for convenience we are dropping the ε in $u(s,\varepsilon)$ etc.):

$$\begin{aligned}
u(s) &= U(s,s^0)u(s^0) + \int_{s^0}^{s} U(s,\sigma)\left[F_y(\sigma)v(\sigma) + R(\sigma,u,v,\varepsilon)\right]d\sigma,\\
v_1(s) &= V_1(s,s^0,\varepsilon)v_1(s^0)\\
&\quad + \varepsilon^{-1}\int_{s^0}^{s} V_1(s,\sigma,\varepsilon)\left[G_{1,\bar{x}}(\sigma)u(\sigma) + G_{1,\bar{y}}(\sigma)\bar{v}(\sigma)\right.\\
&\qquad \left. + S_1(\sigma,u,v,\varepsilon) - \varepsilon Y_1'(\sigma)\right]d\sigma,\\
\bar{v}(s) &= \bar{V}(s,s^0,\varepsilon)\bar{v}(s^0)\\
&\quad + \varepsilon^{-1}\int_{s^0}^{s} \bar{V}(s,\sigma,\varepsilon)\left[\bar{G}_{\bar{x}}(\sigma)u(\sigma) + \bar{G}_{y_1}(\sigma)v_1(\sigma)\right.\\
&\qquad \left. + \bar{S}(\sigma,u,v,\varepsilon) - \varepsilon\bar{Y}'(\sigma)\right]d\sigma.
\end{aligned} \tag{18}$$

We have to provide estimates for all the terms appearing on the right-hand side of (18).

Since all eigenvalues of the matrix $\bar{G}_{\bar{y}}(0)=\bar{A}$ have negative real parts there exists $\mu \in (0,1)$ for $|s^0|$ small enough such that $\bar{G}_{\bar{y}}(s)$ has eigenvalues with real parts smaller than -2μ for $s \in J$. Hence, by a lemma due to Flatto and Levinson [1] we have

$$|\bar{V}(s,\sigma,\varepsilon)| \leqq Ke^{-\mu(s-\sigma)/\varepsilon} \quad \text{for } s^0 \leqq \sigma \leqq s \leqq 0. \tag{19}$$

Using H5 and integrating (17) yields an estimate for V_1:

$$V_1(s,\sigma,\varepsilon) \leqq e^{\hat{k}[(-s)^{\hat{q}} - (-\sigma)^{\hat{q}}]/\varepsilon} \quad \text{for } s^0 \leqq \sigma \leqq s \leqq 0, \tag{20}$$

where $\hat{q} := q+1$, $\hat{k} := k/\hat{q}$.

The fundamental matrix solution $U(s,\sigma)$ satisfies

$$|U(s,\sigma)| \leqq K \quad \text{for } s^0 \leqq \sigma \leqq s \leqq 0. \tag{21}$$

Moreover, we have

$$|F_y(s)| \leqq K, \quad |\bar{G}_{\bar{x}}(s)| \leqq K, \quad |G_{1,\bar{x}}(s)| \leqq K \quad \text{for } s \in J. \tag{22}$$

Remark 4. For $\alpha \geqq 1$, when $\phi(x)$ has continuous partial derivatives, $G_1(x,\phi(x))=0$ for $x \in J \times \bar{\Omega}$ implies $G_{1,x}(x,\phi(x))+G_{1,y}(x,\phi(x))\phi_x(x)=0$, and this together with $G_{1,y}(0,0)=0$ yields $G_{1,x}(0,0)=0$. $G_{1,\bar{x}}(0,0)=0$ is also true for $\alpha<1$ if $G_1(s,\bar{x},y)$ has no linear terms in $\bar{x}$. Hence, in these two cases we will be able to derive a better estimate for $G_{1,\bar{x}}(s)$ than the one given in (22).

Since $G_{1,\bar{y}}(x,y)$ and $\bar{G}_{y_1}(x,y)$ vanish at the origin (see H2) we will also have better estimates for $G_{1,\bar{y}}(s)$ and $\bar{G}_{y_1}(s)$. In order to find these estimates we need the asymptotic behavior of $\bar{X}(s)$ and $\bar{Y}(s)$ as $s \to 0^-$.

In the domain corresponding to the one introduced in (15) F satisfies the estimate

$$|F(s,\bar{x},y_1,\bar{y})| \leqq K\big((-s)+|\bar{x}|+|y_1|+|\bar{y}|\big). \tag{23}$$

From H2 we know that

$$\begin{aligned} G_1(x,y_1,\bar{y}) &= \sum_{i=1}^{m} c_i x_i + a y_1^2 + \hat{G}_1(x,y_1,\bar{y}), \\ \bar{G}(x,y_1,\bar{y}) &= \bar{A}\bar{y} + Bx + \hat{G}(x,y_1,\bar{y}), \end{aligned} \tag{24}$$

where the Taylor formulae of $\hat{G}_1$ and $\hat{G}$ begin with quadratic terms in (x,y) and with a cubic term in y_1. Hence

$$\bar{\phi}(s,\bar{x}) = \bar{w}(s,\bar{x},\phi_1(s,\bar{x})) = \bar{a}s + \tilde{B}\bar{x} + \tilde{G}(s,\bar{x},\phi_1(s,\bar{x})) \tag{25}$$

where the remainder term again is at least quadratic in $(s,\bar{x},\phi_1)$ and cubic in ϕ_1. This together with (10) provides the following estimate

$$|\bar{\phi}(s,\bar{x})| \leqq K\big((-s)^{\tilde{\alpha}}+|\bar{x}|\big) \quad \text{where } \tilde{\alpha} := \min(3\alpha,1). \tag{26}$$

Using (23) and (26) in (12) we find for $\bar{X}(s)$ that for some $\bar{K}>0$

$$|\bar{X}(s)| \leqq \bar{K}\int_s^0 \big((-\sigma)^{\hat{\alpha}}+|\bar{X}(\sigma)|\big)\,d\sigma = \frac{\bar{K}}{\hat{\alpha}+1}(-s)^{\hat{\alpha}+1} + \bar{K}\int_s^0 |\bar{X}(\sigma)|\,d\sigma$$

where $\hat{\alpha} := \min(\alpha,1)$.

Applying the generalized Gronwall lemma we obtain

$$|\bar{X}(s)| \leqq \hat{K}(-s)^{\hat{\alpha}+1}, \qquad s \in J, \tag{27}$$

for some $\hat{K}>0$. From (27) and (10) we get[1]

$$|Y_1(s)| \leqq K_1(-s)^{\alpha_1}, \qquad s \in J, \tag{28}$$

$\alpha_1 := \min(\alpha,2)$, and from (26) and (27)[2]

$$|\bar{Y}(s)| \leqq \tilde{K}(-s)^{\tilde{\alpha}}, \qquad s \in J. \tag{29}$$

---

[1] There is a more precise assertion for $\alpha<1$:

$$Y_1(s) = a_1(-s)^{\alpha} + o\big((-s)^{\alpha}\big), \qquad s \to 0^-.$$

[2] Using (25) directly and footnote 1 we find for $\alpha<1$ that

$$\bar{Y}(s) = \bar{a}s + O\big((-s)^{\tilde{\tilde{\alpha}}}\big), \qquad s \to 0^-,$$

where $\tilde{\tilde{\alpha}} := \min(1+\alpha, 3\alpha)$.

Finally, using (24) this allows us to state

$$\begin{aligned} &|G_{1,\bar{x}}(s)| \leqq C(-s)^{\gamma}, \\ &|G_{1,\bar{y}}(s)| \leqq C(-s)^{\nu}, \qquad s \in J, \\ &|\bar{G}_{y_1}(s)| \leqq C(-s)^{\bar{\alpha}}, \end{aligned} \tag{30}$$

where $\gamma \in \{0, \hat{\alpha}\}$, $\nu = \hat{\alpha}$, $\bar{\alpha} = \min(2\alpha, 1)$ and C some positive constant depending on K.

Remark 5. Usually $\gamma = 0$ as we have seen in (22); however, $\gamma = \hat{\alpha}$ for the two cases indicated in Remark 4.

From (24) we conclude that

$$|\bar{S}(s,u,v,\varepsilon)| \leqq \bar{K}\left(\varepsilon + |u|^2 + |\bar{v}|^2 + (-s)^{\hat{\alpha}}|v_1|^2 + |v_1|^3\right). \tag{31}$$

An analogous estimate holds for S_1 if $G_1(x,y)$ has no quadratic term in y_1. (Of course, all these estimates can be improved if the vector functions F and G are of special form.) Here, we take into account only one such special case corresponding to bifurcation situations similar to the one considered in [4], where $\alpha < 1$ and $G_1(x,y)$ has no linear term in x. Since then, substituting in the first expression of (24) the above estimates for $\bar{X}$, Y_1, $\bar{Y}$, we obtain

$$0 = aa_1(-s)^{2\alpha} + o\left((-s)^{2\alpha}\right), \qquad s \to 0^-,$$

and this implies $a = 0$. Thus

$$|S_1(s,u,v,\varepsilon)| \leqq C_1\left(\varepsilon + |u|^2 + |\bar{v}|^2 + (-s)^{\beta}|v_1|^2 + |v_1|^3\right)$$

where we define

$$\beta := \begin{cases} \alpha & \text{if } \alpha < 1 \text{ and } G_1 \text{ has no linear term in } x, \\ 0 & \text{otherwise.} \end{cases} \tag{32}$$

In order to save writing we will drop the term $|v_1|^3$ of (31) and (32) in what follows, since it can easily be verified that it is always less or equal than the preceding $|v_1|^2$-term in the case of Theorem 2 (cf. (52) and (57)) and also in the case of Theorem 3 (cf. (76)).

The estimates for $|Y_1'(s)|$ and $|\bar{Y}'(s)|$ are still missing. In the case $\alpha \geqq 1$ (H3′) both quantities are bounded. In the case $\alpha < 1$ the first estimate follows from (11) and (27):

$$|Y_1'(s)| \leqq C(-s)^{\beta_1}, \qquad s \in [s^0, 0), \tag{33}$$

where $\beta_1 := \min(\alpha - 1, 0)$. From

$$\bar{Y}'(s) = \frac{d}{ds}\bar{\phi}\left(s, \bar{X}(s)\right) = \frac{d}{ds}\bar{w}\left(s, \bar{X}(s), Y_1(s)\right) = \bar{w}_s(s) + \bar{w}_{\bar{x}}(s)F(s) + \bar{w}_{y_1}(s)Y_1'(s)$$

together with (25) and (33), we conclude

$$|\bar{Y}'(s)| \leqq \bar{C}(-s)^{\bar{\beta}}, \qquad s \in [s^0, 0), \tag{34}$$

where $\bar{\beta} := \min(3\alpha - 1, 0)$.

Now, applying all these estimates to (18) we obtain

(35)

$$|u(s)| \leqq Kc_0\varepsilon + K^2 \int_{s_0}^{s} \Big[|v(\sigma)| + \varepsilon + |u(\sigma)|^2 + |v(\sigma)|^2 \Big] d\sigma,$$

$$|v_1(s)| \leqq c_0\varepsilon + \varepsilon^{-1} \int_{s_0}^{s} e^{\hat{k}[(-s)^{\hat{q}} - (-\sigma)^{\hat{q}}]/\varepsilon} \Big[C(-\sigma)^{\gamma} |u(\sigma)| + C(-\sigma)^{\nu} |\bar{v}(\sigma)| + C_1\Big(\varepsilon + |u(\sigma)|^2 + |\bar{v}(\sigma)|^2 + (-\sigma)^{\beta} |v_1(\sigma)|^2\Big) + C(-\sigma)^{\beta_1}\varepsilon \Big] d\sigma,$$

$$|\bar{v}(s)| \leqq Kc_0\varepsilon + \varepsilon^{-1} K \int_{s_0}^{s} e^{-\mu(s-\sigma)/\varepsilon} \Big[K|u(\sigma)| + C(-\sigma)^{\bar{\alpha}} |v_1(\sigma)| + \bar{K}\Big(\varepsilon + |u(\sigma)|^2 + |\bar{v}(\sigma)|^2 + (-\sigma)^{\hat{\alpha}} |v_1(\sigma)|^2\Big) + \bar{C}(-\sigma)^{\bar{\beta}}\varepsilon \Big] d\sigma.$$

In the subsequent proofs we will also need the following two auxiliary results which are obtained by simple integration:

(36) $$\int_{s_0}^{s} e^{-\mu(s-\sigma)/\varepsilon} d\sigma < \frac{\varepsilon}{\mu},$$

(37) $$\int_{s_0}^{s} e^{\hat{k}[(-s)^{\hat{q}} - (-\sigma)^{\hat{q}}]/\varepsilon} k(-\sigma)^q d\sigma < \varepsilon, \qquad s \in J.$$

Moreover, we are now able to state two lemmas which provide relations between q and α that will prove useful in formulating our results.

LEMMA 1. *q satisfies*

(38) $$q \geqq \hat{\alpha}$$

and if $\beta = \alpha$ (cf. (32))

(39) $$q \geqq 2\alpha.$$

Proof. $G_1(x, y_1, \bar{y})$ is a smooth function with respect to all variables and satisfies $G_1(0,0,0) = 0$, $G_{1,y_1}(0,0,0) = 0$ and $G_{1,\bar{y}}(0,0,0) = 0$ (cf. H2). Hence, G_1 has the following expansion in the neighborhood of the origin

(40) $$G_1(x, y_1, \bar{y}) = L_0(x) + ay_1^2 + L_1(x, \bar{y})\, y_1 + Q_0(x, \bar{y}) + by_1^3 + \hat{G}_1(x, y_1, \bar{y})$$

where L indicates linear terms, Q quadratic terms and the terms in the remainder are at least cubic. This implies

(41) $$G_{1,y_1}(x, y_1, \bar{y}) = 2ay_1 + L_1(x, \bar{y}) + 3by_1^2 + \hat{G}_{1,y_1}(x, y_1, \bar{y})$$

where the terms in the remainder are at least quadratic. The functions $\bar{X}(s)$, $Y_1(s)$ and $\bar{Y}(s)$ satisfy

(42) $$G_1\big(s, \bar{X}(s), Y_1(s), \bar{Y}(s)\big) = 0 \quad \text{for } s \in J.$$

Moreover, H5 implies that there are positive constants k and q such that

(43) $$\big|G_{1,y_1}\big(s, \bar{X}(s), Y_1(s), \bar{Y}(s)\big)\big| \geqq k(-s)^q \quad \text{for } s \in J.$$

Hence, inserting the functions $\bar{X}(s)$, $Y(s)$ in (41) and making use of their asymptotic behavior given in (27), (28) and (29), respectively, we derive that

$$k(-s)^q \leqq k_1(-s)^{\alpha_1} + k_2(-s)^{\tilde{\alpha}}, \qquad s \in J \tag{44}$$

for some positive constants k_1, k_2. And this implies that $q \geqq \min(\alpha_1, \tilde{\alpha})$ which proves (38).

In the case $\beta = \alpha$ we have $\alpha < 1$ and $L_0 \equiv 0$, $a = 0$ in (40). Hence

$$G_1(x, y_1, \bar{y}) = L_1(x, \bar{y})\, y_1 + Q_0(x, \bar{y}) + b y_1^3 + \hat{G}_1(x, y_1, \bar{y}) \tag{45}$$

and

$$G_{1,y_1}(x, y_1, \bar{y}) = L_1(x, \bar{y}) + 3b y_1^2 + \hat{G}_{1,y_1}(x, y_1, \bar{y}). \tag{46}$$

And by the same argument as before we obtain that $q \geqq \min(2\alpha, \tilde{\alpha})$. This implies $q \geqq \bar{\alpha}$ and proves (39) for $\alpha \leqq \frac{1}{2}$.

Assume now that $\alpha > \frac{1}{2}$ and $q < 2\alpha$. Thus, we have $1 \leqq q < 2\alpha < 1 + \alpha < 3\alpha$. Inserting $\bar{X}(s)$, $Y_1(s)$ and $\bar{Y}(s)$ in (46) we find that all terms not in L_1 and all terms in L_1 with $\bar{X}_l(s)$ are $O((-s)^{2\alpha})$. The remaining terms in L_1 are the terms $c_0 s$ and $c_i \bar{Y}_i(s)$, $i = 1, \cdots, n-1$, with the asymptotic behavior $\bar{Y}_i(s) = \bar{a}_i s + O((-s)^{2\alpha})$ as $s \to 0^-$. Thus, (43) can only be satisfied if

$$c_0 + \sum_{i=1}^{n-1} c_i \bar{a}_i \neq 0.$$

Consider now $(45)_s$, i.e. $\bar{X}(s)$, $Y_1(s)$, $\bar{Y}(s)$ inserted in (45). Here, we have the terms $c_0 s Y_1(s)$ and $c_i \bar{Y}_i(s) Y_1(s)$; and all the other terms are $O((-s)^\rho)$, $\rho = \min(2, 3\alpha)$. Hence, (42) admits the representation

$$0 = a_1 c_0 (-s)^{1+\alpha} + a_1 \sum_{i=1}^{n-1} c_i \bar{a}_i (-s)^{1+\alpha} + o\big((-s)^{1+\alpha}\big), \qquad s \to 0^-,$$

implying $c_0 + \Sigma c_i \bar{a}_i = 0$ and leading to a contradiction. □

LEMMA 2. *If $\alpha = q$, then $\alpha \geqq \frac{1}{2}$.*[3]

Proof. We use the same arguments as in the proof of Lemma 1. If $q = \alpha < \frac{1}{2}$, the term $2aY_1(s)$, $a \neq 0$, is the only one in $(41)_s$ that can satisfy (43); all other terms are $O((-s)^{2\alpha})$. And, (42) admits

$$0 = a_1 a (-s)^{2\alpha} + o\big((-s)^{2\alpha}\big), \qquad s \to 0^-.$$

This, however, implies $a = 0$ and, hence, leads to a contradiction. □

4. The main results in the case "$s < 0$". In this section we state the results for the trajectories in the phase-space for the case "$s < 0$". The proofs are given in §5.

Theorem 2 below provides an existence result for the solution $(\bar{x}(s, \varepsilon), y(s, \varepsilon))$ of the initial value problem (5), (6) as well as an estimate for the domain of validity and the error of the approximation $(\bar{X}(s), Y(s))$ which is the solution of the reduced initial value problem (12).

[3] There is strong evidence that $\alpha + q \geqq 1$ also holds in general, under the hypotheses H1–H5.

THEOREM 2. *Suppose the hypotheses* H1–H5 *hold, and let* $\hat{\alpha} := \min(\alpha, 1)$, $\bar{\alpha} := \min(2\alpha, 1)$, $q_1 := 1 - \hat{\alpha} + q$. *If* q *satisfies*

$$q < \min(1 + \gamma, \bar{\alpha} + \nu) \tag{47}$$

where $\gamma \geqq 0$ ($\geqq 1$, *if* $\alpha \geqq 1$) *and* $\nu \geqq \hat{\alpha}$ *are defined by* (30); *then, if* $|s^0|$ *is taken sufficiently small, the following result holds*:

There exist positive constants c *and* $\varepsilon_1 \leqq \varepsilon_0$ *such that the solution* $(\bar{x}(s,\varepsilon), y(s,\varepsilon))$ *of the initial value problem* (5), (6) *exists at least for* $s \in J^* := [s^0, -c\varepsilon^{q^*}]$, *where* $q^* := 1/(q_1 + q - \beta)$, $\beta \in \{0, \alpha\}$ *defined in* (32), *and*

$$|\bar{x}(s,\varepsilon) - \bar{X}(s)| < \begin{cases} c\varepsilon \log(-s)^{-1}, & q_1 = 1, \\ c\varepsilon(-s)^{1-q_1}, & q_1 > 1, \end{cases}$$

$$|y_1(s,\varepsilon) - Y_1(s)| < c\varepsilon(-s)^{-q_1},$$

$$|\bar{y}(s,\varepsilon) - \bar{Y}(s)| < \begin{cases} c\varepsilon \log(-s)^{-1}, & q_1 = 1, \\ c\varepsilon(-s)^{\bar{\alpha}-q_1}, & q_1 > 1 \end{cases}$$

for $s \in J^*$, $\varepsilon \in (0, \varepsilon_1)$.

Remark 6. Since $q \geqq \hat{\alpha}$ (cf. Lemma 1), we always have $q_1 \geqq 1$. There are exactly two cases where $q_1 = 1$: (i) $q = \alpha < 1$ and (ii) $\alpha \geqq 1$, $q = 1$. Note, that we have $q_1 = q$ for $\alpha \geqq 1$, and that always $q^* < 1/q_1 \leqq 1$ (cf. Lemma 1). The restriction (47) on q depends on the lowest order terms of $G_{1,\bar{x}}(s)$ and $G_{1,\bar{y}}(s)$, or in other words is due to the coupling of the y_1-equation to the rest of the system (5). With the assumptions made in this paper we have seen that $\gamma = 0$ or $\gamma = \hat{\alpha}$ and $\nu = \hat{\alpha}$ (see (30) and Remark 5). However, if $G_1(x,y)$ is of special form such that the coupling is weaker, both quantities may be greater than $\hat{\alpha}$ (compare formulas (27)–(30)).

For applications it is often useful to have the estimates of Theorem 2 in a more transparent form with respect to the dependence on ε. We state this slightly weaker result as Theorem 2′. It is obtained from Theorem 2 essentially by considering the right end-point of J^* as being variable, i.e. by varying q^* below its maximum.

THEOREM 2′. *Suppose the same assumptions are satisfied as in Theorem* 2, *and let* $\hat{\alpha} := \min(\alpha, 1)$, $\bar{\alpha} := \min(2\alpha, 1)$, $q_1 := 1 - \hat{\alpha} + q$ *and* $|s^0|$ *be sufficiently small. Then there exist positive constants* c *and* $\varepsilon_1 \leqq \varepsilon_0$ *such that the solution* $(\bar{x}(s,\varepsilon), y(s,\varepsilon))$ *of* (5), (6) *exists at least for* $s \in J^* := [s^0, -c\varepsilon^{\delta_0/q_1}]$, *where* $\delta_0 := q_1/(q_1 + q - \beta)$, $\beta \in \{0, \alpha\}$ *defined by* (32). *Moreover, for every* $\delta \in (0, \delta_0]$ *we have*

$$|\bar{x}(s,\varepsilon) - \bar{X}(s)| < \begin{cases} c\varepsilon \log \varepsilon^{-1}, & q_1 = 1, \\ c\varepsilon^{1-\delta+\delta/q_1}, & q_1 > 1, \end{cases}$$

$$|y_1(s,\varepsilon) - Y_1(s)| < c\varepsilon^{1-\delta},$$

$$|\bar{y}(s,\varepsilon) - \bar{Y}(s)| < \begin{cases} c\varepsilon \log \varepsilon^{-1}, & q_1 = 1, \\ c\varepsilon^{1-\delta+\bar{\alpha}\delta/q_1}, & q_1 > 1 \end{cases}$$

for $s \in [s^0, -c\varepsilon^{\delta/q_1}]$, $\varepsilon \in (0, \varepsilon_1)$.

Remark 7. Note that, the smaller the s-interval in which the approximation $(\bar{X}(s), Y(s))$ is considered the better the error estimates. Only in the case $q_1 = 1$ the estimates for the $\bar{x}$- and $\bar{y}$-components are independent of δ.

5. Proofs. In this section we give the proofs of the theorems stated in §4.

Proof of Theorem 2. (a) We first consider the case $\alpha<1$. Define

$$
\begin{aligned}
&s^*(\varepsilon):=-k^*\varepsilon^{q^*},\\
&r(s,\varepsilon):=\begin{cases}k_0\varepsilon\log(-s)^{-1}, & q_1=1,\\ k_0\varepsilon(-s)^{1-q_1}, & q_1>1,\end{cases}\\
&r_1(s,\varepsilon):=k_1\varepsilon(-s)^{-q_1},\\
&\bar r(s,\varepsilon):=\begin{cases}\bar k_0\varepsilon\log(-s)^{-1}, & q_1=1,\\ \bar k_0\varepsilon(-s)^{\bar\alpha-q_1}, & q_1>1,\end{cases}
\end{aligned}
\tag{48}
$$

where all the constants involved are supposed to be positive, $q_1\geqq 1$, $q^*q_1<1$, and will be specified more precisely later. These functions are defined and continuous for $\varepsilon\in[0,\varepsilon_0]$ and $s\in[s^0,0)$. Moreover, the positive functions r, r_1, $\bar r$ increase as s increases.

Now, consider the set $\mathfrak{d}\subset\mathbb{R}^{m+n}$ defined by

$$
\mathfrak{d}:=\{|u|<r\}\times\{|v_1|<r_1\}\times\{|\bar v|<\bar r\}\times\{s\in J^*\}
\tag{49}
$$

where $J^*:=[s^0,s^*]$. For ε small enough, $\mathfrak{d}$ lies in the domain considered.

With respect to $\mathfrak{d}$ the solution $(u(s,\varepsilon),\ v(s,\varepsilon))$ of the initial value problem (14), (16) is assumed to exist on $[s^0,m_+)$, where m_+ may depend on ε.

For ε and $|s^0|$ sufficiently small we can always achieve that

$$
r(s)\leqq\bar r(s)<r_1(s)<1\quad\text{for } s\in J^*.
\tag{50}
$$

Thus, from the estimates given in (35) and also using (36), (37) we derive the following estimates for the solution of (14), (16):

$$
\begin{aligned}
&|u(s)|\leqq M\Big[\varepsilon+\int_{s^0}^{s}r_1(\sigma)\,d\sigma\Big],\\
&|v_1(s)|\leqq M\big[\varepsilon(-s)^{-q}+(-\tau)^{\gamma-q}r(\tau)+(-\tau)^{\nu-q}\bar r(\tau)\\
&\qquad\qquad+\bar r^2(s)(-s)^{-q}+(-s)^{\beta-q}r_1^2(s)+(-\tau)^{\beta_1-q}\varepsilon\big],\\
&|\bar v(s)|\leqq M\big[\varepsilon+r(s)+(-\tau)^{\bar\alpha}r_1(\tau)+\bar r^2(s)+(-s)^{\hat\alpha}r_1^2(s)+(-\tau)^{\bar\beta}\varepsilon\big],
\end{aligned}
\tag{51}
$$

for $s\in[s^0,m_+)$, where M is a positive constant depending on all constants appearing in (35) and on θ_1 introduced in (15), and τ stands for s or s^0, respectively, depending on whether the exponent of the corresponding expression is negative or nonnegative (e.g. $(-\tau)^{\bar\alpha}r_1(\tau)=k_1\varepsilon(-\tau)^{\bar\alpha-q_1}\to k_1\varepsilon(-s)^{\bar\alpha-q_1}$ if $q_1>\bar\alpha$). Since $\alpha<1$, we have $\hat\alpha=\alpha$ and $\beta_1=\alpha-1<0$, $\bar\beta=\min(3\alpha-1,0)$ and $\gamma\geqq 0$, $\nu\geqq\alpha$. We may write (51) more explicitly (for simplicity we write down only the case $q_1>1$):

$$
\begin{aligned}
&|u(s)|\leqq M\Big[\varepsilon+k_1\varepsilon\int_{s^0}^{s}(-\sigma)^{-q_1}d\sigma\Big],\\
&|v_1(s)|\leqq M\big[\varepsilon(-s)^{-q}+k_0\varepsilon(-\tau)^{1-q_1+\gamma-q}+\bar k_0\varepsilon(-\tau)^{\bar\alpha-q_1+\nu-q}\\
&\qquad\qquad+\bar k_0^2\varepsilon^2(-s)^{2\bar\alpha-2q_1-q}+k_1^2\varepsilon^2(-s)^{\beta-2q_1-q}+\varepsilon(-s)^{\alpha-1-q}\big],\\
&|\bar v(s)|\leqq M\big[\varepsilon+k_0\varepsilon(-s)^{1-q_1}+k_1\varepsilon(-s)^{\bar\alpha-q_1}\\
&\qquad\qquad+k_1^2\varepsilon^2(-s)^{\alpha-2q_1}+\bar k_0^2\varepsilon^2(-s)^{2\bar\alpha-2q_1}+\varepsilon(-\tau)^{\bar\beta}\big],\qquad\text{for } s\in[s^0,m_+).
\end{aligned}
\tag{52}
$$

We now choose $q_1 := 1-\alpha+q$. (By Lemma 1 we know that $q_1 \geqq 1$.) We want to show that

$$
\begin{aligned}
&|u(s)| \leqq \frac{r}{2} \\
&|v_1(s)| \leqq \frac{r_1}{2} \qquad \text{for } s \in [s^0, m_+). \\
&|\bar{v}(s)| \leqq \frac{\bar{r}}{2}
\end{aligned} \tag{53}
$$

We only have to consider the case $\tau = s$ since for $\tau = s^0$ the right-hand sides of (52) are always smaller (for the same choice of constants) than in the case $\tau = s$.

Let us consider the case $q_1 > 1$ first. It can easily be seen, by taking into account the assertions of Lemma 1 and Lemma 2, that if q satisfies

$$q < 1+\gamma \quad \text{and} \quad q < \bar{\alpha}+\nu \tag{54}$$

then, for $|s^0|$ and ε sufficiently small, (53) holds under the following conditions concerning the constants involved:

$$M\left(1+\frac{k_1}{q_1-1}\right) \leqq \frac{k_0}{2}, \quad M(6) \leqq \frac{k_1}{2}, \quad M(4+k_0+k_1) \leqq \frac{\bar{k}_0}{2} \tag{55}$$

and

$$k^* := k_1^2, \qquad q^* := \frac{1}{q_1+q-\beta}. \tag{56}$$

It is obvious that (55) can be satisfied by choosing $k_1 (\geq 1)$ first (so that the second inequality holds), k_0 second and then $\bar{k}_0$ large enough.

In the case $q_1 = 1$, where $\alpha = q$ and $\bar{\alpha} = 1$ (by means of Lemma 2), and $\beta = 0$ (cf. (39), Lemma 1), the situation is exactly the same only with k_1 having a factor 1 instead of $1/(q_1-1)$ in the first inequality of (55). (54) is automatically satisfied. Having shown (53), the proof of Theorem 2, as far as the case $\alpha < 1$ is concerned, is completed by applying the Global Existence Theorem for ode's; and by choosing c as the maximum of k_0, k_1, $\bar{k}_0$ and k^*, and ε_1 sufficiently small.

(b) The case $\alpha \geqq 1$ is formally completely analogous and we therefore only point out the differences in the formulas. We have $\hat{\alpha} = 1$, $\beta_1 = 0$, $\bar{\beta} = 0$, $\beta = 0$ and $\gamma \geqq 1$, $\nu \geqq 1$, $\bar{\alpha} = 1$. Thus, for $q_1 > 1$, we obtain instead of (52)

$$
\begin{aligned}
&|u(s)| \leqq M\left[\varepsilon + k_1\varepsilon \int_{s^0}^{s} (-\sigma)^{-q_1} d\sigma\right], \\
&|v_1(s)| \leqq M\left[\varepsilon(-s)^{-q} + k_0\varepsilon(-\tau)^{1-q_1+\gamma-q}\right. \\
&\qquad\qquad \left. + \bar{k}_0\varepsilon(-\tau)^{1-q_1+\nu-q} + k_1^2\varepsilon^2(-s)^{-2q_1-q}\right], \\
&|\bar{v}(s)| \leqq M\left[\varepsilon + k_0\varepsilon(-s)^{1-q_1} + k_1\varepsilon(-s)^{1-q_1} + k_1^2\varepsilon^2(-s)^{1-2q_1} + \bar{k}_0^2\varepsilon^2(-s)^{2-2q_1}\right],
\end{aligned} \tag{57}
$$

for $s \in [s^0, m_+)$, where M again is a common constant.

We choose $q_1 := q$ ($\geqq 1$, by Lemma 1); and we find, first for $q_1 > 1$, that if

$$q < 1+\gamma \quad \text{and} \quad q < 1+\nu \tag{58}$$

then, for $|s^0|$ and ε sufficiently small, the estimate (53) holds under the following conditions:

$$M\left(1+\frac{k_1}{q_1-1}\right) \leqq \frac{k_0}{2}, \quad M(4) \leqq \frac{k_1}{2}, \quad M(3+k_0+k_1) \leqq \frac{\bar{k}_0}{2} \tag{59}$$

and

$$k^* := k_1^2, \qquad q^* := \frac{1}{q_1+q}. \tag{60}$$

The inequalities (59) can again be satisfied by choosing k_1 ($\geqq 1$) first such that the second inequality holds and then taking k_0, $\bar{k}_0$ large enough.

In the case $q_1 = 1$, where (58) holds anyway, again the only difference lies in the factor 1 of k_1 in the first inequality of (59).

The proof is completed in the same way as in part (a). □

Proof of Theorem 2'. Since the functions $r(s,\varepsilon)$, $r_1(s,\varepsilon)$ and $\bar{r}(s,\varepsilon)$, defined in (48), are increasing functions of s, Theorem 2 implies that

$$\begin{aligned}
&|\bar{x}(s,\varepsilon) - \bar{X}(s)| < \begin{cases} k_0 q^* \varepsilon \log \varepsilon^{-1}, & q_1 = 1, \\ k_0 k^{*-(q_1-1)} \varepsilon^{1-(q_1-1)q^*}, & q_1 > 1, \end{cases} \\
&|y_1(s,\varepsilon) - Y_1(s)| < k_1 k^{*-q_1} \varepsilon^{1-q_1 q^*}, \\
&|\bar{y}(s,\varepsilon) - \bar{Y}(s)| < \begin{cases} \bar{k}_0 q^* \varepsilon \log \varepsilon^{-1}, & q_1 = 1, \\ \bar{k}_0 k^{*-(q_1-\bar{\alpha})} \varepsilon^{1-(q_1-\bar{\alpha})q^*}, & q_1 > 1, \end{cases}
\end{aligned} \tag{61}$$

for $s \in [s^0, -k^* \varepsilon^{q^*}]$, $\varepsilon \in (0, \varepsilon_1)$, where we have replaced the common constant c by the original constants. Now, let $\delta \in (0, \delta_0]$, where $\delta_0 := q_1/(q_1+q-\beta)$. Then, (61) also holds for $q^* = \delta/q_1$. Thus, $\delta = \delta_0$ gives the maximal q^* defined in Theorem 2, $\delta < \delta_0$ means considering a smaller s-interval on which we have better estimates, however, due to Theorem 2. If we take only q^* variable (by varying δ) but k^* fixed in $s^*(\varepsilon) = -k^* \varepsilon^{q^*}$, then we can take the same constants c and ε_1 as in Theorem 2, since $q^* < 1$, $k^* \geqq 1$, and they are independent of δ. □

6. The case "$s > 0$". In this section we will deal with the case where the flow along a stable branch of the reduced manifold of (1) leads away from the point at which the stability assumption (4) is violated. This situation is encountered in bifurcation diagrams of the type sketched in Fig. 1. Another possible situation is given in Fig. 3.

In our formulation of the problem with H1 supposed to hold this second case will occur for $s > 0$ in the orbit equations (5). We again want to provide a maximal domain of validity (with sharpest possible error estimates) of an approximation given by a reduced trajectory. And again we have to consider only a small neighborhood of that unstable point where Theorem 1 is not yet applicable ("not yet" instead of "no more" as in the case "$s < 0$").

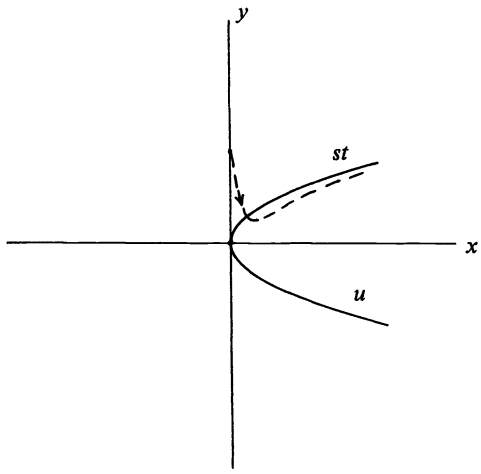


FIG. 3

The reason why this case is different from the one considered in the previous sections and in fact more involved is that the initial conditions to the orbit equations (5) can no longer be assumed, by means of Theorem 1, to satisfy a condition H4 with an estimate ε. And since we want to give a general result for this case we do not know the estimates in the transition interval[4] which depend on the global problem at hand. Hence, we know neither the starting point $s=s^*(\varepsilon)>0$ for the approximation nor the approximate size of possible initial conditions. The following questions, however, seem to be appropriate: What is the smallest possible $s^*(\varepsilon)$, and what estimates for the initial conditions are required in order that the reduced trajectory considered approximates the corresponding trajectory of the full system on the interval $[s^*(\varepsilon),s^1]$ for some $s^1\in(0,1)$, independent of ε? And what is the sharpest possible estimate that can be achieved in a given situation? Theorem 3 below will give an answer to these questions.

Let us become more precise now. Besides assumption H1 we again suppose that H2 is satisfied. Let us denote the yet unknown initial conditions to the system (5) by

$$\bar{x}(s^*,\varepsilon)=\bar{x}^0(\varepsilon),\qquad y(s^*,\varepsilon)=y^0(\varepsilon) \tag{62}$$

where $s^*(\varepsilon)>0$ for $\varepsilon\in(0,\varepsilon_0)$ and $s^*\to 0$ as $\varepsilon\to 0$.

The appropriate s-interval in H3 is now $\hat{J}:=[0,s^1]$ for some $0<s^1<1$, independent of ε; and in the formulas (10), (11) $(-s)$ has to be replaced by s. Let us denote the so adjusted hypothesis by $\hat{\text{H}}3$, and let us assume that the hypotheses $\hat{\text{H}}3'$ and $\hat{\text{H}}5$ which have to be altered in the analogous way hold, too. The assumption H4 has to be changed into the following.

$\hat{\text{H}}4$. The solution $(\bar{X}(s), Y(s))$ of (12), where $Y(s):=\phi(s,\bar{X}(s))$, exists for $s\in\hat{J}$.

Under these hypotheses H1, H2, $\hat{\text{H}}3$–$\hat{\text{H}}5$ estimates for the solution corresponding to (62) of the system (14) can be derived completely analogous to those given in §3. We

[4] That is, the ε-dependent interval containing the unstable point (and overlapping with the interval J^* of Theorem 2) where the solution of the IVP (5), (6) cannot be approximated by a solution of the reduced system $(5)_{\varepsilon=0}$ (cf. [6] and Part II of this paper).

only state the final assertion which corresponds to (35):

$$|u(s)| \leqq K|u(s^*)| + K^2 \int_{s^*}^{s} \left[|v(\sigma)| + \varepsilon + |u(\sigma)|^2 + |v(\sigma)|^2 \right] d\sigma,$$

$$\begin{aligned} |v_1(s)| &\leqq e^{-\hat{k}(s^{\hat{q}} - s^{*\hat{q}})/\varepsilon} |v_1(s^*)| \\ &\quad + \varepsilon^{-1} \int_{s^*}^{s} e^{-\hat{k}(s^{\hat{q}} - \sigma^{\hat{q}})/\varepsilon} \Big[C\sigma^{\gamma} |u(\sigma)| + C\sigma^{\nu} |\bar{v}(\sigma)| \\ &\quad + C_1 \left(\varepsilon + |u(\sigma)|^2 + |\bar{v}(\sigma)|^2 + \sigma^{\beta} |v_1(\sigma)|^2 \right) + C\sigma^{\beta_1} \varepsilon \Big] d\sigma, \end{aligned} \tag{63}$$

$$\begin{aligned} |\bar{v}(s)| &\leqq K e^{-\mu(s-s^*)/\varepsilon} |\bar{v}(s^*)| \\ &\quad + \varepsilon^{-1} K \int_{s^*}^{s} e^{-\mu(s-\sigma)/\varepsilon} \Big[K|u(\sigma)| + C\sigma^{\bar{\alpha}} |v_1(\sigma)| \\ &\quad + \bar{K} \left(\varepsilon + |u(\sigma)|^2 + |\bar{v}(\sigma)|^2 + \sigma^{\hat{\alpha}} |v_1(\sigma)|^2 \right) + \bar{C}\sigma^{\bar{\beta}} \varepsilon \Big] d\sigma \end{aligned}$$

for $s \in [s^*, s^1]$, with μ and $\hat{q}$, $\hat{k}$ defined in (19) and (20), respectively. And again estimates analogous to (36) and (37) hold:

$$\int_{s^*}^{s} e^{-\mu(s-\sigma)/\varepsilon} d\sigma < \frac{\varepsilon}{\mu}, \tag{64}$$

$$\int_{s^*}^{s} e^{-\hat{k}(s^{\hat{q}} - \sigma^{\hat{q}})/\varepsilon} k\sigma^{q} d\sigma < \varepsilon, \qquad s \in [s^*, s^1]. \tag{65}$$

The estimate (65) can also be replaced by one which is sometimes better for our purposes:

$$\int_{s^*}^{s} e^{-\hat{k}(s^{\hat{q}} - \sigma^{\hat{q}})/\varepsilon} d\sigma \leqq \int_{s^*}^{s} e^{-\psi(s)(s-\sigma)/\varepsilon} d\sigma < \frac{\varepsilon}{\psi(s)} \tag{66}$$

where $\psi(s) := \hat{k} s^q$.

We will also need the following estimate

$$\int_{s^*}^{s} e^{-\hat{\psi}(s-\sigma)/\varepsilon} \sigma^{\rho} d\sigma < \frac{\varepsilon}{\hat{\psi}} s^{\rho} + \hat{\rho} \frac{\varepsilon^2}{\hat{\psi}^2} s^{*\rho - 1}, \qquad s \in [s^*, s^1] \tag{67}$$

where $\rho \in \mathbb{R}$, $\hat{\psi} = \psi(s)$ or $\hat{\psi} = \mu$, respectively, and

$$\hat{\rho} = \begin{cases} 0, & \rho \geqq 0, \\ -\rho, & \rho < 0. \end{cases}$$

The two lemmas of §3 hold true, of course, also in the case "$s > 0$". In order to be able to prove Theorem 3 we need one more estimate which we state as

LEMMA 3. *For every $\rho \geqq 0$ we have*

$$e^{-\hat{k}(s^{\hat{q}} - \sigma^{\hat{q}})/\varepsilon} \leqq \left(\frac{\sigma}{s} \right)^{\rho}$$

if $(\rho\varepsilon/k)^{1/\hat{q}} \leqq \sigma \leqq s$.

Proof. The function $h(\tau) := e^{-\hat{k}\tau^{\hat{q}}/\varepsilon\tau^{\rho}}$ is decreasing for $\tau > (\rho\varepsilon/k)^{1/\hat{q}}$. □

Remark 8. Lemma 3 holds, of course, also for the special choice $\hat{k} = \mu$ and $\hat{q} = 1$ with $k = \mu$.

We now state the main result for the case "$s > 0$". It again provides an existence statement for a solution $(\bar{x}(s,\varepsilon),\ y(s,\varepsilon))$ of the orbit equations (5) but now to the initial conditions (62), as well as a maximal domain of validity with sharpest possible error estimates for the approximation $(\bar{X}(s),\ Y(s))$ which is the solution of the reduced initial value problem (12). Theorem 3, in contrast to Theorem 2, contains two free parameters due to the fact that the size of the initial values is unknown.

THEOREM 3. *Suppose the hypotheses* H1, H2, Ĥ3–Ĥ5 *hold, and let* $\hat{\alpha} := \min(\alpha, 1)$, $\bar{\alpha} := \min(2\alpha, 1)$ *and let* $\beta \in \{0, \alpha\}$ *be defined by* (32). *Moreover, assume that* q_1 *and* $\bar{\omega}$ *are in* $[0,1)$ *and such that*

$$\beta \leqq \bar{\omega} \leqq q_1 \leqq \min(\bar{\alpha}, 2q)$$

and let

$$q^* := \frac{1}{1+2q-\hat{\alpha}-\beta}, \qquad \delta := (q_1+q-\beta)q^*. \tag{68}$$

If q satisfies

$$q \leqq q_1+\gamma \quad \text{and} \quad q \leqq \bar{\omega}+\nu \tag{69}$$

(*for* $\bar{\omega} = \bar{\alpha}$ *the equality sign does not apply in the second inequality*) *where* $\gamma \geqq 0$ ($\geqq 1$, *if* $\alpha \geqq 1$) *and* $\nu \geqq \hat{\alpha}$ *are defined by* (30), *then the following result holds*:

There exist positive constants c, $\hat{s}^1 \leqq s^1$ *and* $\varepsilon_1 \leqq \varepsilon_0$ *such that the solution* $(\bar{x}(s,\varepsilon),\ y(s,\varepsilon))$ *of the initial value problem* (5), (62) *exists for* $s \in \hat{J}^* := [s^*(\varepsilon), \hat{s}^1]$, *where* $s^*(\varepsilon) := c\varepsilon^{q^*}$, *provided the initial values at* $s = s^*$ *satisfy*

$$\begin{aligned} &|\bar{x}^0(\varepsilon) - \bar{X}(s^*)| = O(\varepsilon^{\delta}), \\ &|y_1^0(\varepsilon) - Y_1(s^*)| = o(\varepsilon^{\delta_1}), \qquad \delta_1 := (q-\beta)q^*, \\ &|\bar{y}^0(\varepsilon) - \bar{Y}(s^*)| = o(\varepsilon^{\bar{\delta}}), \qquad \bar{\delta} := (\bar{\omega}+q-\beta)q^*, \end{aligned} \tag{70}$$

where the little o can be replaced by capital O in the y_1*-component if* $q_1 = 0$ *and in the* $\bar{y}$*-component if* $\bar{\omega} = q_1$.

Moreover, $(\bar{x}(s,\varepsilon),\ y(s,\varepsilon))$ *possesses the following estimates*

$$\begin{aligned} &|\bar{x}(s,\varepsilon) - \bar{X}(s)| < c\varepsilon^{\delta}, \\ &|y_1(s,\varepsilon) - Y_1(s)| < c\varepsilon^{\delta} s^{-q_1}, \\ &|\bar{y}(s,\varepsilon) - \bar{Y}(s)| < c\varepsilon^{\delta} s^{\bar{\omega}-q_1}, \end{aligned} \tag{71}$$

for $s \in \hat{J}^*$ *and* $\varepsilon \in (0, \varepsilon_1)$, $(\bar{X}(s),\ Y(s))$ *being the solution of the reduced initial value problem* (12).

Remark 9. Note that $0 < \delta < 1$, and that $\delta_1 \leqq \bar{\delta} \leqq \delta$, and in particular $\delta_1 < \delta$, except if $q_1 = 0$ (which is usually excluded by (69)). This allows to have less rigid requirements for the initial value $y_1^0(\varepsilon)$ (or also for $\bar{y}^0(\varepsilon)$ if $\bar{\omega}$ is chosen in an appropriate way),

which is an important fact since these initial values are, in general, provided by weaker estimates in the transition interval. Applying Theorem 3, one has to choose the two parameters q_1 and $\bar{\omega}$ such that the given initial values satisfy (70) and the q, ν, γ at hand satisfy (69). q_1 is, of course, taken as large as possible since this increases δ and yields optimal estimates (71) (or the other way around: taking the optimal δ by means of (70) gives a possible q_1 defined by (68), if (69) as well as $q_1<1$ and $\leqq \min(\bar{\alpha}, 2q)$[5] is satisfied). Note also that q^*, and hence the interval of validity $\hat{J}^*$ of the approximation $(\bar{X}(s), Y(s))$, is independent of q_1, $\bar{\omega}$ and δ. Moreover, $q^*=1/2q$ and $\delta=(q_1+q)/2q$ for $\alpha \geqq 1$.

There again exists the slightly weaker version of Theorem 3 with the estimates (71) not depending on s but only on ε. We are not going to state this Theorem 3′ here, since it can be derived from Theorem 3 in a way completely analogous to how Theorem 2′ has been obtained from Theorem 2.

7. Proof of Theorem 3. The proof is analogous to the one for Theorem 2. We define

$$
\begin{aligned}
&s^*(\varepsilon) := k^*\varepsilon^{q^*},\\
&r(\varepsilon) := k_0\varepsilon^{\delta}, \quad r_1(s,\varepsilon) := k_1\varepsilon^{\delta}s^{-q_1}, \quad \bar{r}(s,\varepsilon) := \bar{k}_0\varepsilon^{\delta}s^{\bar{\omega}-q_1}
\end{aligned} \tag{72}
$$

where $k^*(\geqq 1)$, k_0, k_1, $\bar{k}_0$ are positive constants, and $0 \leqq q_1 < 1$, $0 \leqq \bar{\omega} \leqq q_1$, $0<\delta<1$ and $q_1 q^* < \delta$. All these quantities will be specified more precisely later. The positive functions r, r_1, $\bar{r}$ are defined and continuous for $\varepsilon \in (0, \varepsilon_0)$ and $s \in (0, s^1]$ and decrease as s increases (if they are not constant).

Furthermore, we suppose that the initial values (62) to the system (5) obey the estimates

$$
\begin{aligned}
&|\bar{x}(s^*,\varepsilon) - \bar{X}(s^*)| \leqq c_0\varepsilon^{\lambda},\\
&|y_1(s^*,\varepsilon) - Y_1(s^*)| \leqq c_1\varepsilon^{\lambda_1},\\
&|\bar{y}(s^*,\varepsilon) - \bar{Y}(s^*)| \leqq \bar{c}_0\varepsilon^{\bar{\lambda}},
\end{aligned} \tag{73}
$$

where c_0, c_1, $\bar{c}_0$ are positive quantities and λ, λ_1, $\bar{\lambda}$ are positive constants $\leqq 1$.

Now, consider the set $\mathfrak{d} \subset \mathbb{R}^{m+n}$ defined by

$$
\mathfrak{d} := \{|u|<r\} \times \{|v_1|<r_1\} \times \{|\bar{v}|<\bar{r}\} \times \{s \in \hat{J}^*\} \tag{74}
$$

where $\hat{J}^* := [s^*(\varepsilon), \hat{s}^1]$, $\hat{s}^1$ independent of ε and $0<\hat{s}^1 \leqq s^1$. For ε and $\hat{s}^1$ small enough, $\mathfrak{d}$ lies in the domain considered.

With respect to $\mathfrak{d}$ the solution $(u(s,\varepsilon), v(s,\varepsilon))$ of the initial value problem (14), (73) is assumed to exist on $[s^*, m_+)$.

For ε and $\hat{s}^1$ small enough we can always achieve that

$$
r \leqq \bar{r}(s) \leqq r_1(s) < 1 \quad \text{for } s \in \hat{J}^*. \tag{75}
$$

[5] Since $2q \geqq \hat{\alpha} + q$ (cf. Lemma 1), there is strong evidence that always $2q \geqq 1$ (cf. footnote 3) and therefore $q_1 \leqq 2q$ is automatically satisfied.

(a) We consider first the case $\alpha<1$. Thus, from (63) and also using (64)–(67) we derive the following estimates for the solution $(u(s,\varepsilon), v(s,\varepsilon))$:

$$
\begin{aligned}
|u(s)| &\leqq M\left[c_0\varepsilon^{\lambda}+k_1\varepsilon^{\delta}\int_{s^*}^{s}\sigma^{-q_1}\,d\sigma\right],\\
|v_1(s)| &\leqq M\Big[e^{-\hat{k}(s^{\hat{q}}-s^{*\hat{q}})/\varepsilon}c_1\varepsilon^{\lambda_1}+\varepsilon s^{-q}+k_0\varepsilon^{\delta}s^{\gamma-q}\\
&\qquad +\bar{k}_0\varepsilon^{\delta}\left(s^{\nu+\bar{\omega}-q_1-q}+\varepsilon\hat{h}(\nu+\bar{\omega}-q_1)s^{*\nu+\bar{\omega}-q_1-1}s^{-2q}\right)\\
&\qquad +k_1^2\varepsilon^{2\delta}\left(s^{\beta-2q_1-q}+\varepsilon\hat{h}(\beta-2q_1)s^{*\beta-2q_1-1}s^{-2q}\right)\\
&\qquad +\varepsilon\left(s^{\alpha-1-q}+\varepsilon s^{*\alpha-2}s^{-2q}\right)\Big],\\
|\bar{v}(s)| &\leqq M\Big[e^{-\mu(s-s^*)/\varepsilon}\bar{c}_0\varepsilon^{\bar{\lambda}}+\varepsilon+k_0\varepsilon^{\delta}+k_1\varepsilon^{\delta}s^{\bar{\alpha}-q_1}\\
&\qquad +k_1^2\varepsilon^{2\delta}\left(s^{\alpha-2q_1}+\varepsilon\hat{h}(\alpha-2q_1)s^{*\alpha-2q_1-1}\right)\\
&\qquad +\bar{k}_0^2\varepsilon^{2\delta}\left(s^{2\bar{\omega}-2q_1}+\varepsilon\hat{h}(\bar{\omega}-q_1)s^{*2\bar{\omega}-2q_1-1}\right)+\varepsilon\left(s^{\bar{\beta}}+\varepsilon\hat{h}(\bar{\beta})s^{*\bar{\beta}-1}\right)\Big]
\end{aligned}
\tag{76}
$$

for $s\in[s^*,m_+)$, where M is a positive constant depending on all constants appearing in (63), and on s^1 (but not on $\hat{s}^1$),

$$
\hat{h}(z):=\begin{cases}0, & z\geqq 0,\\ -z, & z<0.\end{cases}
$$

We want to show that

$$
\begin{aligned}
|u(s)| &\leqq \frac{r}{2},\\
|v_1(s)| &\leqq \frac{r_1}{2} \qquad \text{for } s\in[s^*,m_+).\\
|\bar{v}(s)| &\leqq \frac{\bar{r}}{2}
\end{aligned}
\tag{77}
$$

For that, we have to get rid of the two exponential terms in (76). This can be done by applying Lemma 3 for $\sigma=s^*$, $\rho=q_1$ and $\rho=q_1-\bar{\omega}$, respectively, and yields the following two estimates

$$
\begin{aligned}
e^{-\hat{k}(s^{\hat{q}}-s^{*\hat{q}})/\varepsilon}c_1\varepsilon^{\lambda_1} &\leqq c_1k^{*q_1}\varepsilon^{\lambda_1+q_1q^*}s^{-q_1},\\
e^{-\mu(s-s^*)/\varepsilon}\bar{c}_0\varepsilon^{\bar{\lambda}} &\leqq \bar{c}_0k^{*(q_1-\bar{\omega})}\varepsilon^{\bar{\lambda}+(q_1-\bar{\omega})q^*}s^{\bar{\omega}-q_1},
\end{aligned}
\tag{78}
$$

provided

$$
q^*\leqq\frac{1}{1+q} \quad\text{and}\quad k^*\geqq\max\left(\left(\frac{q_1}{k}\right)^{1/(1+q)},\ \frac{q_1-\bar{\omega}}{\mu}\right)
\tag{79}
$$

holds.

It can easily be seen, by taking into account the assertions of Lemma 1, that if

$$
q_1\leqq\min(\bar{\alpha},2q), \qquad q\leqq q_1+\gamma \quad\text{and}\quad q\leqq\bar{\omega}+\nu
\tag{80}
$$

then, for $\hat{s}^1$ and ε sufficiently small, (77) holds under the following conditions concerning the constants involved

$$
\begin{aligned}
&M(c_0+1)\leqq\frac{k_0}{2},\\
&M\big(1+c_1+k_0+\bar{k}_0+6\big)\leqq\frac{k_1}{2},\\
&M(1+\bar{c}_0+k_0+8)\leqq\frac{\bar{k}_0}{2},
\end{aligned}
\tag{81}
$$

and

$$
\begin{aligned}
&q^*:=\frac{1}{1+2q-\alpha-\beta},\qquad \delta:=(q_1+q-\beta)q^*\\
&\beta\leqq\bar{\omega}<\bar{\alpha},\\
&\lambda\geqq\delta,\quad \lambda_1\geqq(q-\beta)q^*,\quad \bar{\lambda}\geqq(\bar{\omega}+q-\beta)q^*
\end{aligned}
\tag{82}
$$

as well as

$$
\begin{aligned}
&k^*:=\max\left(1,k_1^{2/(q_1+q-\beta)},\ k_1^2,\ \bar{k}_0^2,\ \left(\frac{q_1}{k}\right)^{1/(1+q)},\ \frac{q_1-\bar{\omega}}{\mu}\right)\\
&c_1\leqq k^{*-q_1},\ \bar{c}_0\leqq k^{*\bar{\omega}-q_1}\quad\text{(only if } q_1>0 \text{ or } \bar{\omega}<q_1\text{, respectively)}.
\end{aligned}
\tag{83}
$$

It is obvious that (81) can be satisfied by choosing k_0 first (such that the first inequality holds), $\bar{k}_0$ second and finally k_1 large enough.

We have excluded $\bar{\omega}=\bar{\alpha}$ in (82). In this case everything holds true, too, if $q<\bar{\alpha}+\nu$. We only get slightly different inequalities (81), namely $\bar{k}_0$ in the second inequality has to be replaced by 1 and k_1 has to be added to the expression in parentheses of the third inequality. And the only effect is that k_1, instead of $\bar{k}_0$, has to be chosen second after k_0, in order to satisfy the three inequalities.

Having shown (77) the proof of Theorem 3, as far as the case $\alpha<1$ is concerned, is completed by applying the Global Existence Theorem for ode's; and by taking ε_1 sufficiently small and choosing c as the maximum of k_0, k_1, $\bar{k}_0$ and k^*.

(b) The case $\alpha\geqq 1$ is formally completely analogous to the case $\alpha<1$. Since $\beta=\bar{\beta}=\beta_1=0$ the last term in the right-hand side of the v_1-component of (76) as well as the last term in the right-hand side of the $\bar{v}$-component can be dropped. Moreover, α and $\bar{\alpha}$ have to be replaced by 1 (hence $\bar{\omega}=\bar{\alpha}$ is not possible), and we know that $\gamma\geqq 1$ and $\nu\geqq 1$ in this case. The rest is identical to part (a). □

Acknowledgment. The author wishes to thank Professor R. E. O'Malley, Jr. for the invitation to visit the Mathematical Sciences Department of Rensselaer Polytechnic Institute, Troy, New York, and for providing such a good working atmosphere during this year.

REFERENCES

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

[2] F. Hoppensteadt, *Properties of solutions of ordinary differential equations with small parameters*, Comm. Pure Appl. Math., 24 (1971), pp. 807–839.

[3] N. R. Lebovitz and R. J. Schaar, *Exchange of stabilities in autonomous systems*, Stud. Appl. Math., 54, 3 (1975), pp. 229–260.

[4] _____, *Exchange of stabilities in autonomous systems—II. Vertical bifurcation*, Stud. Appl. Math., 56 (1977), pp. 1–50.

[5] E. F. MISHCHENKO AND N. KH. ROZOV, *Differential Equations with Small Parameters and Relaxation Oscillations*, Plenum Press, New York, 1980.

[6] K. NIPP, *An algorithmic approach to singular perturbation problems in ordinary differential equations with an application to the Belousov-Zhabotinskii reaction*, Ph.D. thesis, ETH Zurich, No. 6643, 1980.

[7] _____, *An extension of Tikhonov's theorem in singular perturbations for the planar case*, Z. Angew. Math. Phys., 34 (1983), pp. 277–290.

[8] L. S. PONTRYAGIN, *Asymptotic properties of solutions of systems of differential equations with a small parameter multiplying leading derivatives*, Izv. AN SSSR, Ser. Mat., 21, 5 (1957), pp. 605–626; AMS Transl. Ser. 2, 18 (1961), pp. 295–319.

[9] L. S. PONTRYAGIN AND E. F. MISHCHENKO, *Asymptotic behavior of the solutions of systems of differential equations with a small parameter multiplying the higher derivatives*, Izv. AN SSSR, Ser. Mat., 23, 5 (1959), pp. 643–660. (In Russian.)

[10] W. WASOW, *Asymptotic Expansions for Ordinary Differential Equations*, John Wiley, New York, 1965.

SIAM J. MATH. ANAL.
Vol. 17, No. 3, May 1986

003

SAUTS DES SOLUTIONS DES ÉQUATIONS $\varepsilon\ddot{x}=f(t,x,\dot{x})$*

FRANCINE DIENER†

Abstract. One of the main difficulties in the study of the equations $\varepsilon\ddot{x}=f(t,x,\dot{x})$ for small ε is created by the existence of parts of solutions, called "jumps" for which the velocity is large. We give here precise definitions of such jumps, their extremities, and their thickness, and we show that, for most equations of the type considered, it is easy to compute, up to an infinitesimal, the position of jumps, extremities and thickness for each solution. Our methods use both nonstandard analysis and a geometrical approach consisting of, among other things, observing solutions in some convenient planes, (plans d'observabilité) of the phase space.

Nous proposons ci-dessous l'étude des sauts des solutions des équations différentielles du type

$$\text{(I)} \qquad \varepsilon\ddot{x}=f(t,x,\dot{x}), \qquad \varepsilon \text{ infinitesimal,}^{1} \text{ positif,}$$

c'est-à-dire l'étude des portions de solutions parcourues à vitesse grande (*). On appelle souvent ces sauts des "couches" limites, libres, ou intérieures. On sait que la présence de sauts est liée à la petitesse de ε et qu'elle constitue l'une des principales difficultés dans l'étude du comportement des solutions de ces équations.

Nous montrons que, pour une classe étendue d'équations de ce type, on peut, moyennant un choix convenable d'échelle de vitesse, décrire simultanément les sauts de *toutes* les solutions d'une équation donnée, calculer leur équation à un infinitésimal près, et déterminer pour chacun d'eux son extrémité en fonction de son origine, ainsi que le temps nécessaire pour le parcourir.

Les équations du type envisagé ont intéressé, au cours des trente dernières années, de nombreux auteurs qui se sont principalement attachés à prouver l'existence d'*une* solution d'un problème aux limites (et plus rarement d"un problème de conditions initiales) qui présente certains types de sauts à des instants prescrits. Des techniques plus ou moins élaborées de développements asymptotiques, faisant intervenir un changement d'échelle du temps t (là où nous utilisons un changement d'échelle de vitesse), permettent de traiter un tel problème aux limites dans les cas semi-linéaires (f ne dépend pas de $\dot{x}$) et quasi-linéaires (f est linéaire en $\dot{x}$) [24], [19], [9]. Dans les cas où f n'est pas linéaire en $\dot{x}$, les études sont beaucoup plus rares [23]. Les résultats les plus importants obtenus à ce jour sont dus à Howes [12], [11], qui à l'aide de méthodes d'inéquations différentielles, retrouve facilement et étend les résultats de ses prédécesseurs dans le cas où f est un polynôme en $\dot{x}$ de degré 0, 1, ou 2. Quelques articles proposent également d'astucieuses généralisarions à certains cas où f est "superquadratique" en $\dot{x}$ [14].

Au premier chapitre nous donnons des définitions précises des objets étudiés (sauts, leur origine, leur extrémité, leur épaisseur), définitions absentes des études précédentes et qui apparaissent très naturellement ici grâce au point de vue non

* Received by the editors October 25,, 1983, and in revised form December 15, 1984.

† Institut de Mathématiques de l'Université d'Oran, B.P. 1524, Es-Senia, ORAN, Algérie.

[1] Nous adoptons dans cet article le point de vue non standard qui, comme cela a été souvent développé [20], [16], présente bien des avantages dans ce contexte. Pour guider un lecteur qui ignorerait ces nouvelles "règles du jeu", nous proposons en annexe un index terminologique (auquel renvoie le symbole (*) placé après le mot concerné) comprenant les principaux résultats non standard dont nous avons besoin et, entre crochets, les équivalents anglosaxons [Robinson/Nelson] lorsque ceux-ci existent.

standard adopté. Nous précisons également la classe d'équations étudiée. On verra qu'elle est très générale et contient, en particulier, les cas usuels, semi-linéaires, quasi-linéaires, et quadratiques. Nous présentons au second chapitre la méthode employée sur un exemple simple. Enfin, nous indiquons au troisième chapitre les principaux résultats obtenus et donnons leur démonstration.

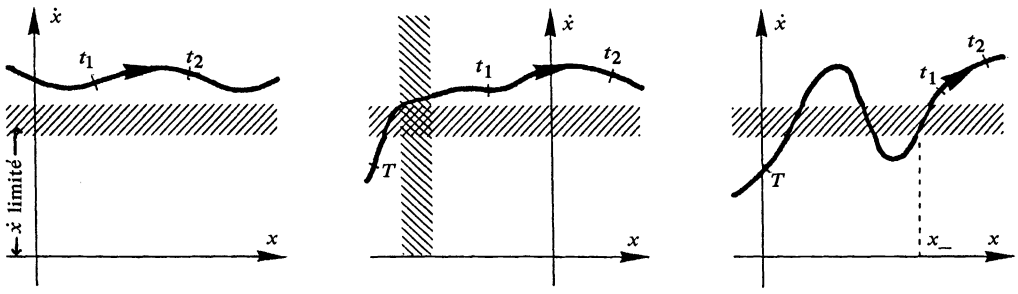
1. Quelques définitions.

Sauts d'une fonction. Soit $x: I \to \mathbb{R}$ une fonction interne (*) de classe C^1. On dit que $x(t)$ présente un *saut* sur l'intervalle $[t_1, t_2] \subset I$ si, sur cet intervalle, la vitesse $\dot{x}(t)$ est grande et si $x(t_1) \not\approx x(t_2)$ (*), avec $x(t_1)$ et $x(t_2)$ limités (*). On notera que l'intervalle $[t_1, t_2]$ est nécessairement petit (*), et que, comme la variation de x sur cet intervalle est supposée appréciable (*), les sauts apparaissent sur le graphe $(t, x(t))$ de la fonction comme des portions de ce graphe d'ombre (*) verticale, d'où leur nom. Une fonction qui présente un saut sur l'intervalle $[t_1, t_2]$ n'est pas S-continue (*) sur cet intervalle.

Comme, par continuité, $\dot{x}(t)$ ne peut changer de signe sans devenir limitée, tout saut est soit croissant, soit décroissant. Sauf mention contraire, on se bornera désormais au cas des sauts croissants, celui des sauts décroissants s'en déduisant immédiatement.

Pour t_0 réel standard (*), on dit que la fonction x présente un saut à l'instant t_0 s'il existe un intervalle $[t_1, t_2] \subset I$ contenu dans le halo de t_0 (*) sur lequel x présente un saut. On notera que $\dot{x}$ n'est pas nécessairement grand, ni même défini au point t_0.

Origine et extrémité d'un saut. Si $x(t)$ est une fonction présentant un saut sur l'intervalle $[t_1, t_2]$, on peut, dans certains cas, définir l'origine de ce saut; on peut, biensûr, procéder de manière analogue pour définir l'extrémité d'un saut. Trois cas sont à envisager (Fig. 1):



Saut sans origine *Saut d'origine $-\infty$* *Saut d'origine x_-*

FIG. 1

1) Ou bien, pour tout $t \in I$, si $t \leqq t_1$, alors $\dot{x}(t)$ est grand. On dit que *le saut sur* $[t_1, t_2]$ *est sans origine sur* I.

2) Ou bien il existe $T \in I$, $T < t_1$, tel que $\dot{x}(T)$ est limité, mais pour tout $t \in [T, t_1]$, si $x(t)$ est limité alors $\dot{x}(t)$ est grand. On dit que *le saut sur* $[t_1, t_2]$ *a son origine en* $x = -\infty$ (ou $+\infty$ dans le cas d'une fonction décroissante). On notera que ceci ne signifie pas que x tend vers $-\infty$, mais seulement que x "n'arrête pas de sauter" avant d'atteindre (à reculons) des réels grands négatifs.

3) Ou bien il existe $T \in I$, $T < t_1$, tel que $x(T)$ et $\dot{x}(T)$ sont limités, et $\dot{x}(t)$ est positif pour tout $t \in [T, t_2]$; x est donc croissante sur cet intervalle. Dans ce cas on souhaite définir l'*origine du saut sur* $[t_1, t_2]$ comme *le standard* (x_-) *tel que, pour tout* $t \in [T, t_1]$, *si* $x(t) \gg x_-$ (*), *alors* $\dot{x}(t)$ *est grand, et au contraire, si* $x(t) \ll x_-$, *alors il*

existe $\tau \in [t, t_1]$ *tel que* $\dot{x}(\tau)$ *est limité.* Notons que, comme les notions de "grand" et "limité" sont externes, il serait illusoire d'espérer pouvoir remplacer les "doubles chevrons" $\gg$ ou $\ll$ par de simples inégalités (internes) $>$ ou $<$.

Pour établir l'existence de x_-, on procède de la manière suivante: considérons tout d'abord la coupure externe (C_-, C_+) de $\mathbb{R}$ définie par $C_- = \{\xi \in \mathbb{R}$, tels qu'il existe $\tau \in [T, t_1]$, avec $\dot{x}(\tau)$ limité et $x(\tau) \geqq \xi\}$ et $C_+ = \mathbb{R} \setminus C_-$.

Remarquons que $x(T) \in C_-$, $x(t_1) \in C_+$, et $x(T)$ ainsi que $x(t_1)$ sont limités. On pose alors $S_- = {}^S C_- =$ standardisé $(*)$ de C_-, c'est à dire l'unique sous-ensemble standard de $\mathbb{R}$ ayant les mêmes éléments standards que C_-; de façon analogue, on pose $S_+ = {}^S C_+$. A présent, comme tout élément de C_+ domine tout élément de C_-, c'est en particulier vrai pour tout élément standard: ceci montre que tout élément standard de S_+ domine tout élément standard de S_-, d'où, par transfert, S_+ domine S_-. On démontre de même que (S_+, S_-) est une partition de $\mathbb{R}$. Comme ${}^0x(T) - 1 \in S_-$, $S_- \neq \varnothing$; et comme ${}^0x(t_1) + 1 \in S_+$, $S_+ \neq \varnothing$; (S_-, S_+) constitute donc une coupure standard de $\mathbb{R}$: elle détermine un unique standard x_-.

Vérifions que le standard x_- a bien la propriété souhaitée. Soit $t \in [T, t_1]$ tel que $x(t) \ll x_-$. Comme x est croissante, on a par ailleurs que $x(T) \leqq x(t) \leqq x(t_1)$; $x(t)$ est donc limité; soit $x_0 = {}^0x(t)$. Par hypothèse $x_0 < x_-$; donc $(x_0 + x_-)/2 \in S_-$, et comme il est standard, $(x_0 + x_-)/2 \in C_-$. Il existe donc $\tau \in [T, t_1]$ tel que $\dot{x}(\tau)$ est limité et $(x_0 + x_-)/2 \leqq x(\tau)$. A présent $x(t) \simeq x_0 \ll (x_0 + x_-)/2 \leqq x(\tau)$. D'où $x(t) < x(\tau)$; comme x est croissante, ceci montre que $\tau \in [t, t_1]$. On montre tout aussi facilement que pour $t \in [T, t_1]$, so $x(t) \gg x_-$ alors $\dot{x}(t)$ est grand. CQFD

Notons que d'après le principe de Fehrele $(*)$, non seulement les $x(t) \gg x_-$ sont tels que $\dot{x}(t)$ est grand, mais il existe $t_- \in [T, t_1]$ tel que $x(t_-) \simeq x_-$, et que $\dot{x}(t)$ est grand pour tout $t \geqq t_-$ (inférieur à t_2). On ne peut espérer un résultat analogue pour les $x(t) \ll x_-$, comme le montre l'exemple de la Fig. 2.

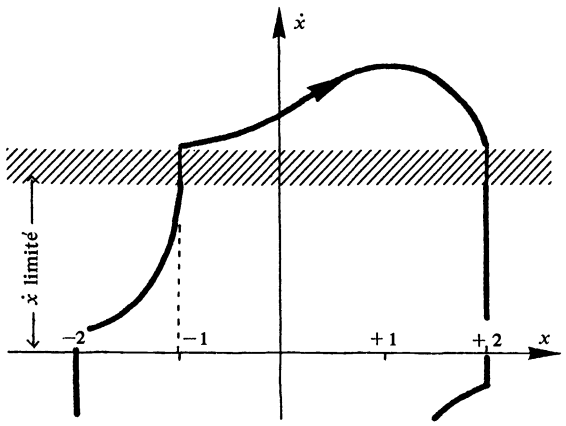


FIG. 2. *La solution périodique de l'équation de Van der Pol* $\varepsilon\ddot{x} + (x^2 - 1)\dot{x} + x = 0$ *présente un saut d'origine* -1 *et d'extrémité* 2. *La vitesse est limitée tant que* $x(t) \ll -1$ *et grande dès que* $x(t) \simeq -1$.

Si l'on considère cette fois le problème de l'extrémité d'un saut, on sera amené à distinguer comme ci-dessus, parmi les sauts (croissants), ceux qui n'ont pas d'extrémité, ceux qui ont $+\infty$ pour extrémité, et ceux qui ont une extrémité x_+ ayant la propriété analogue de celle déterminant l'origine x_- du saut. Lorsqu'un saut de x possède à la fois une origine et une extrémite non infinie, on dira qu'il est *borné*. A l'inverse, si l'une ou l'autre est infinie, ou n'existe pas, on dira qu'il est *non borné*.

Epaisseur d'un saut. Outre l'origine et l'extrémité, une autre notion est utile pour la caractérisation d'un saut: c'est son "épaisseur", qui indique, grosso modo, le temps nécessaire pour accomplir le saut. Voici la définition que nous adopterons [3], généralisant celle de [16].

Nous supposerons systématiquement que la fonction $x: I \to \mathbb{R}$ considérée est monotone et qu'elle présente un saut à l'instant t_0 ayant au moins une extrémité finie. Nous supposerons en outre que ce saut est croissant: on déduit facilement la définition correspondante dans le cas d'un saut décroissant.

Considérons tout d'abord le cas d'un saut borné, d'origine x_- et d'extrémité x_+. On appelle *épaisseur du saut* l'ensemble

$$\mathscr{E} = \{ t \in I, \text{ tels que } x_- \ll x(t) \ll x_+ \}.$$

On vérifie facilement qu'il s'agit d'un ensemble strictement externe (*), et plus précisement d'une galaxie (*) contenue dans le halo de t_0 (voir exemples ci-dessous). Il apparait donc qu'une bonne connaissance des galaxies de $\mathbb{R}$ (voir l'annexe à ce sujet) sera utile pour étudier les épaisseurs de sauts.

Considérons à présent le cas d'un saut non borné, et supposons qu'il n'a pas d'origine ou qu'il a $-\infty$ pour origine. Soit x_+ son extrémité. L'épaisseur du saut sera alors l'ensemble

$$\mathscr{E} = \{ t \in I, \text{ tels que } x(t) \ll x_+ \}.$$

C'est également une galaxie, mais elle n'est pas nécessairement petite; elle contient tout un segment initial de I.

Cas des solutions des équations (I). Soit f une fonction interne, définie sur $\mathbb{R}^3$ et suffisamment régulière pour que l'équation différentielle

$$\text{(I)} \qquad \varepsilon \ddot{x} = f(t, x, \dot{x})$$

possède la propriété d'existence locale et d'unicité des solutions pour $\varepsilon > 0$. De n'imposer à f que le fait d'être interne permet de considérer non seulement les cas où f est standard, mais aussi les cas où $f = F(t, x, \dot{x}, \varepsilon)$, et plus généralement les cas où $f(t, x, \dot{x}) = F(t, x, \dot{x}, a)$, où F est standard et a est un paramètre pouvant prendre des valeurs non standard.

Considérons le champ lent-rapide (*) associé à l'équation (I) dans l'espace des phases $(t, x, \dot{x} = v)$:

$$\text{(II)} \qquad \begin{aligned} \dot{t} &= 1, \\ \dot{x} &= v, \\ \varepsilon \dot{v} &= f(t, x, v). \end{aligned}$$

La trajectoire $(t, x(t), v(t))$ de (II) correspondant à une solution $x(t)$ de l'équation (I) qui présente un saut sur l'intervalle $[t_1, t_2]$ reste entièrement contenue, pendant ce laps de temps, dans l'une des parties H^+ ou H^- de $\mathbb{R}^3$ où v est grand (H^+ si le saut est croissant, H^- s'il est décroissant). Elle reste également proche, durant ce laps de temps, du plan vertical d'équation $t = t_0$, où $t_0 = {}^0t_1$. Notons que les parties H^+, H^-, et le complémentaire de $H^+ \cup H^-$ noté G, sont strictement externes. Lorsqu'une trajectoire de (II), correspondant à une solution présentant un saut est de coordonnées limitées en un instant T antérieur au saut, il découle de ce qui précède que le saut possède une

origine x_-. Nous dirons que "*la trajectoire sort de la région G* (des vitesses limitées) *avec une abscisse* x_-". De même, si le saut a une extrémité x_+, nous dirons que "*la trajectoire rentre dans G avec une abscisse* x_+"

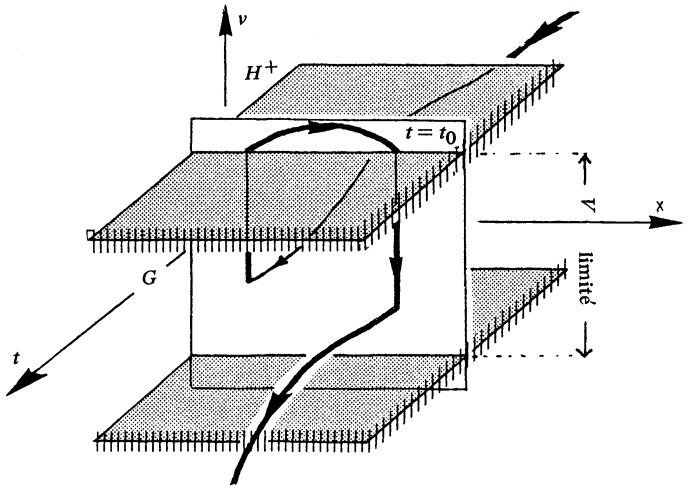


FIG. 3. *Trajectoire du champ* (II) *associé à une solution de* (I) *présentant un saut croissant à l'instant* t_0.

Voici quelques exemples de sauts de solutions et leur épaisseur:

1°) La solution $x(t)=th(t/\varepsilon)$ de l'équation $\varepsilon\ddot{x}=-2x\dot{x}$ présente en $t_0=0$ un saut d'origine $x_-=-1$ et d'extrémité $x_+=1$. Son épaisseur est l'ensemble des réels pouvant s'écrire $t=\varepsilon s$, où s est limité, ensemble qu'on appelle la $\varepsilon-$galaxie de 0(*).

2°) La solution $x(t)=\sqrt{\varepsilon}/(t+\sqrt{\varepsilon})$, $t>-\sqrt{\varepsilon}$, de l'équation $\varepsilon\ddot{x}=2x^3$ présente un saut non borné pour $t_0=0$, d'extrémité $x_+=0$. Son épaisseur est la réunion de $]-\sqrt{\varepsilon}, 0]$ et de la moitié positive de la galaxie des réels pouvant s'écrire $t=\sqrt{\varepsilon}\,s$, où s est limité, encore appelée $\sqrt{\varepsilon}$ −galaxie de 0.

Dans ces deux exemples, comme dans bon nombre d'exemples classiques, les épaisseurs sont réunion d'une partie interne (éventuellement vide) et d'une galaxie très simple (la ε on $\sqrt{\varepsilon}$ −galaxie par example) dite galaxie linéaire (*). Cela se traduit en pratique par le fait qu'il est possible "d'étaler" ces sauts au moyen d'une homothétie. Ainsi les changements de temps $s=t/\varepsilon$ et $s=t/\sqrt{\varepsilon}$ transforment respectivement les examples précédents en $th(s)$ et $1/(s+1)$, fonctions qui ne présentent plus de saut. Cette propriété, que possèdent beaucoup de sauts, de pouvoir s'étaler au moyen d'un changement de temps linéaire $s=(t-t_1)/\varphi(\varepsilon)$, est en fait, à la base de plusieurs méthodes d'étude des sauts, en particulier la méthode des "développements asymptotiques recollés". Cependant, il existe également de nombreuses équations (I) dont les solutions présentent des sauts n'ayant pas cette propriété. En termes d'épaisseur, cela se traduit par le fait que $\mathscr{E}$ s'exprime à l'aide de galaxies non linéaires (*). C'est l'existence de sauts d'épaisseur non linéaire qui conduit à dépasser l'idée un peu trop naïve de caractériser l'épaisseur par un *nombre* ou "ordre de grandeur" (ε ou $\sqrt{\varepsilon}$ par exemple), et de lui substituer un *ensemble* $\mathscr{E}$ strictement externe. Voici un exemple de saut d'épaisseur non linéaire.

3°) La solution $x(t)=-\varepsilon\log(t+e^{-1/\varepsilon})$, $t>-e^{-1/\varepsilon}$, de l'équation $\varepsilon\ddot{x}=\dot{x}^2$ présente un saut non borné en $t=0$, d'extrémité $x_+=0$. Un calcul simple montre que son épaisseur est la réunion de $]-e^{-1/\varepsilon}, 0]$ et de la galaxie des réels pouvant s'ecrire

$t=e^{-1/\varepsilon s}$, avec $s>0$ limité, qu'on appelle la (moitié positive de la) ε–micro-galaxie (*). On notera que cette galaxie contient par exemples les deux nombres $e^{-1/\varepsilon}$ et $e^{-1/2\varepsilon}$ qui ne sont pas du "même ordre de grandeur" puisque leur rapport n'est pas appréciable.

Type de croissance de f en ẋ. Voici à présent quelques définitions concernant le comportement de la fonction f lorsque $\dot{x}$ est grand, qui vont nous permettre de préciser quelle classe d'équations (I) nous nous proposons d'étudier. Nous nous bornerons au cas des vitesses positives, celui des vitesses négatives s'en déduisant aisément.

Soit $F(v)$ une fonction interne de classe $S^0(*)$, continûment dérivable, définie pour tout v positif grand, non décroissante, positive et non petite. Comme F est interne, il existe nécessairement v_0 standard tel que F soit définie et possède encore les propriétés ci-dessus pour tout $v\in[v_0, +\infty[$. On dira que F est le *type de croissance de f par rapport à v pour les v positifs*, s'il existe deux fonctions internes $a(t,x)$ et $r(t,x,v)$ continues et de classe S^0 telles que, pour tout t et x limités, on ait

$$f(t,x,v)=a(t,x)F(v)+r(t,x,v)$$

et

$$r(t,x,v)/F(v)\simeq 0$$

pour tout v positif grand. On appellera la fonction $a(t,x)$ la *mantisse* de f (pour les v positifs). On remarquera que si F et $\bar{F}$ sont deux types de croissance de f, il existe un réel positif R tel que pour tout v grand, $F(v)/\bar{F}(v)\simeq R$. En particulier, si F et $\bar{F}$ sont standard, elles diffèrent par une constante multiplicative pour tout v grand, donc, par le principe de Cauchy (*), pour tout $v\geqq v_0$, avec v_0 standard. Par contre le choix de F est parfaitement arbitraire sur $[0, v_0[$, v_0 standard.

On notera également que si l'on se restreint aux fonctions f et F standard, il est équivalent de supposer que f possède la fonction F comme type de croissance par rapport à v ou que, pour tout t et x limités, $f(t,x,\cdot)$ est un "0" de F et que la limite, quand v tend vers $+\infty$, de f/F existe.

Voici quelques exemples de fonctions avec ou sans type de croissance:

1°) Les cas classiques où f est quasi-linéaire ($f=a(t,x)$), semi-linéaire ($f=a(t,x)v+b(t,x)$), et quadratique ($f=a(t,x)v^2+b(t,x)v+c(t,x)$) ont pour type de croissance respectivement $F(v)=1$, $F(v)=v$, et $F(v)=v^2$, et pour mantisse les fonctions $a(t,x)$.

2°) Plus généralement, si f est un polynôme de degré n en v, ou une somme de "monômes" du type $a(t,x)v^\nu$ où ν est réel positif, le type de croissance de f par rapport à v est v^n, ou v^{ν_0}, où ν_0 est l'exposant le plus grand de v, et la mantisse est le "coefficient" de v^{ν_0}.

3°) Dans l'exemple suivant (étudié par Howes [13])

$$\varepsilon\ddot{x}=(x-t^2)^{2q+1}(1+\dot{x}^2)^{3/2},$$

le type de croissance est $F(v)=v^3$ et la mantisse $(x-t^2)^{2q+1}$.

4°) Dans l'exemple suivant (étudié par Levinson [15])

$$\varepsilon\ddot{x}=4\dot{x}/(3-\dot{x}^4)-x,$$

le type de croissance est $F(v)=1$ (-1 dans le cas des vitesses négatives) et la mantisse est $-x$ (on trouvera une étude nonstandard de cet exemple dans [22]).

5°) Dans les exemples précédents, les types de croissance sont toujours des puissances de v. On peut cependant fort bien envisager d'autres types de croissance. Ainsi, par exemple:

$$\varepsilon\ddot{x}=a(t,x)\dot{x}\log\dot{x} \quad \text{ou} \quad \varepsilon\ddot{x}=a(t,x)th\,\dot{x}\cdots.$$

6°) On construit facilement des exemples d'équations (I) pour lesquelles f n'a pas de type de croissance par rapport à $\dot{x}$ (indépendent de t et x), comme par exemple $\varepsilon\ddot{x}=\sin t\dot{x}$ ou $\varepsilon\ddot{x}=e^{x\dot{x}}$. Cependant, il ne semble pas que de telles équations aient fait l'objet d'études jusqu'ici (l'auteur serait intéressée de connaître une telle étude, pour autant qu'elle existe, ou à défaut une application qui la motiverait).

Les résultats principaux de cet article concernent les équations (I) pour lesquelles f a un type de croissance par rapport à $\dot{x}$. Il convient en outre de supposer que ce type de croissance ne croit pas "trop" vite, et plus précisément, que

$$\text{(III)} \qquad \int_{v_0}^{+\infty}\frac{v\,dv}{F(v)}=+\infty.$$

Cette condition est équivalente, dans le cas des fonctions puissance, au fait que le type de croissance F est au plus quadratique ($\nu_0\leqq 2$). On notera cependant qu'il existe des fonctions "sur-quadratiques", par exemple $F(v)=v^2\log v$, qui la vérifient également. Nous posons cette hypothèse car lorsque le type de croissance de f est plus grand (par exemple $F(v)=v^3$), aucune solution de l'équation ne présente plus de saut, à l'exception éventuellement de ce qu'il est convenu d'appeler des sauts singuliers, c'est à dire des sauts tels que la trajectoire de (II) associée, soit contenue dans le halo de la surface d'équation $f(t,x,v)=0$ (la démonstration de ce fait, ainsi que l'étude des sauts singuliers fait l'objet d'une étude séparée [7]).

2. Un exemple. Dans ce chapitre nous présentons sur un exemple simple la plupart des résultats qui seront démontrés dans le cas général au chapitre suivant. Nous avons choisi pour cela la famille d'équations

$$\text{(IV)} \qquad \varepsilon\ddot{x}=x\dot{x}^{[s]}, \qquad 0<s\leqq 2$$

où $\dot{x}^{[s]}$ est défini par $\dot{x}^{[s]}=|\dot{x}|^{s-1}\dot{x}$.

Outre sa grande simplicité, cette famille présente l'avantage suivant: les portions de solutions ne contenant pas de saut étant toutes presque-constantes (*), les seules portions non triviales des solutions sont des sauts, qui précisément nous intéressent ici.

L'équation correspondant à $s=1$ (qui est intégrable explicitement) est souvent utilisée pour illustrer des sauts de solutions, en particulier dans le cadre des problèmes aux limites [19], [12], [16]. Comme le point de vue adopté ici permet de traiter sans plus de difficulté les cas où $0<s\leqq 2$, nous nous plaçons d'emblée dans ce cadre un peu plus général; ceci nous permet en outre de mettre en évidence, dans un cas particulier, le lien qui existe entre le type de croissance de l'équation, ici $F(v)=v^{[s]}$, et les propriétés des sauts.

Comportement des solutions à vitesse limitée. Ce n'est pas notre propos ici d'étudier les propriétés des solutions lorsque leur vitesse est limitée; cependant, il sera plus facile de comprendre le rôle joué par les sauts si l'on connait le comportement global des solutions et donc, en particulier, leur comportement à vitesse limitée.

Considérons le champ lent-rapide associé à (IV).

$$\text{(V)} \qquad \begin{aligned} &\dot{t}=1,\\ &\dot{x}=v,\\ &\varepsilon\dot{v}=xv^{[s]}, \qquad 0<s\leqq 2. \end{aligned}$$

Bien que ce champ de vecteurs ne vérifie pas, aux points $v=0$, la condition de Lipschitz envlorsque $s<1$, on notera cependant que, par un tel point (distinct de $(t,0,0)$), il passe une unique trajectoire, qui est le graphe d'une solution de l'équation

$$\frac{dx}{dv}=\frac{\varepsilon v}{xv^{[s]}}$$

qui vérifie, elle, la condition de Lipschitz en x. On a donc unicité des solutions de ce champ.

Du fait que ε est supposé petit, il est facile de décrire, à un infinitésimal près, le comportement des trajectoires du champ dans la région G de l'espace où v est limité (Fig. 4). Par les raisonnements usuels de champs lents-rapides (voir par exemple [16]), on montre que, le long d'une trajectoire, aucune variation appréciable de x n'est possible tant que v (et x) sont limités.

Notons également que le comportement des trajectoires du champ (V) est le même quelque soit $s\in]0,2]$.

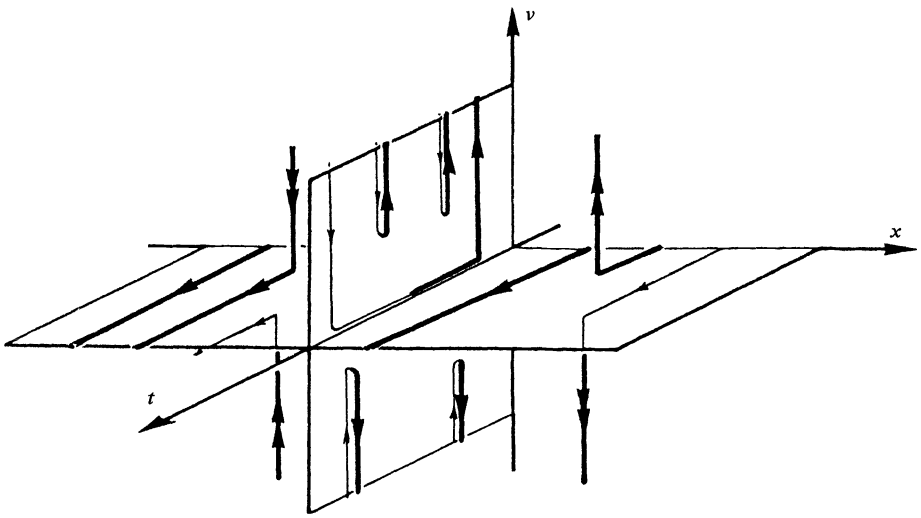


FIG. 4. *Les trajectoires du champ associé à l'équation* $\varepsilon\ddot{x}=x\dot{x}^{[s]}$ *dans l'espace des phases sont presque verticales tant qu'elles restent hors des halos des plans d'équation* $v=0$ *et* $x=0$. *Dans le halo du premier, elles sont proches de droites "horizontales", x constant, et dans le halo du second, elles sont proches de "demi-droites verticales", t constant, ou de "U".*

Comportement des solutions à vitesse grande. En dehors de la région G, c'est à dire dans l'une des régions H^+ ou H^- des points d'ordonnée v grande, positive ou négatives respectivement, la troisième composante du champ (V) cesse d'être grande par rapport aux autres; les trajectoires n'ont plus le comportement caricatural, typique des champs

lents-rapides, évoqué ci-dessus. Une étude particulière est nécessaire. Il n'est pas facile de la mener dans l'espace des phase: les portions de trajectoires sont "trop éloignée". Afin de les "rapprocher", on utilise un macroscope (*), qui est, ici, un changement d'échelle (*) de vitesse. Ce macroscope devra évidement être d'autant plus puissant que la vitesse sur les trajectoires considérées est grande. Cette vitesse dépend, comme on l'imagine, du type de croissance de l'équation et donc, ici, de s. On verra au paragraphe suivant qu'il convient de choisir, lorsque $F(v)=v^{[s]}$, le macroscope suivant:

$$\text{(VI)} \qquad v=\begin{cases}(((2-s)/\varepsilon)V)^{[1/(2-s)]} & \text{si } s\neq 2,\\ \sigma\exp(\sigma V/\varepsilon) & \text{si } s=2,\end{cases}$$

où σ vaut $+1$ ou -1 selon que V est positif ou négatif. A la nouvelle échelle, le champ (V) s'écrit

$$\begin{aligned}&\dot t=1,\\ &\dot x=v,\\ &\dot V=xv \quad \text{où } v \text{ est donné par (VI).}\end{aligned}$$

Nous nous proposons de décrire, à un infinitésimal près, les portions de trajectoires de ce champ, contenues dans le région H_ε^+, image de H^+ par le macroscope, et celles contenues dans la région H_ε^-, image de H^-. Notons que H_ε^+ et H_ε^- contiennent, en particulier, tous les points qui, à cette échelle, ont une ordonnée V non petite, positive ou négative, respectivement.

On ne change pas les trajectoires, pour $V\neq 0$, en multipliant le champ précédent par la fonction nulle part nulle $1/v$ (où $v=v(V)$ est défini par (VI)). cette constatation élémentaire révèle le point de vue "trajectorien" (voir par example [10]) adopté ici, qui fait l'originalité de notre méthode. Il consiste à envisager les trajectoires plutôt comme des objets géométriques de $\mathbb{R}^3$ que comme des graphes de fonctions solutions. De ce fait, le paramétrage étant indifferent, on peut le changer de façon que les sauts, initialement parcourus à vitesse grande, le soit alors à vitesse limitee. La quantité $1/v$ est petite sur H^+ et H^-. Le champ $\mathscr{C}$ ainsi obtenu est donc équivalent, sur $H_\varepsilon^+\cup H_\varepsilon^-$, à un champ $\mathscr{C}_0$ standard et intégrable

$$\mathscr{C}\ \begin{cases}t'=1/v,\\ x'=1,\\ V'=x,\end{cases}\quad\simeq\quad \mathscr{C}_0\ \begin{cases}t'=0,\\ x'=1,\\ V'=x.\end{cases}$$

Les trajectoires de $\mathscr{C}_0$ (Fig. 5) sont des paraboles d'équation

$$t=t_0,\quad V=x^2/2+V_0,\quad t_0,V_0\in\mathbb{R}.$$

On notera que $\mathscr{C}_0$ est indépendant de s. En d'autres termes, si le macroscope dépend, lui, de s, ce que l'on "voit" sous le macroscope, au contraire, est identique, à un petit près, pour toutes les valeurs de s considérées.

Sauts des solutions des équations (IV). Ceci va permettre d'étudier les sauts des solutions des équations (IV). Soit $x(t)$ une solution de (IV) présentant un saut sur $[t_1,t_2]$. La trajectoire correspondante dans l'espace des phase reste contenue dans H (si le saut est croissant) ou H^- (sinon); sous le macroscope, cette trajectoire est le graphe

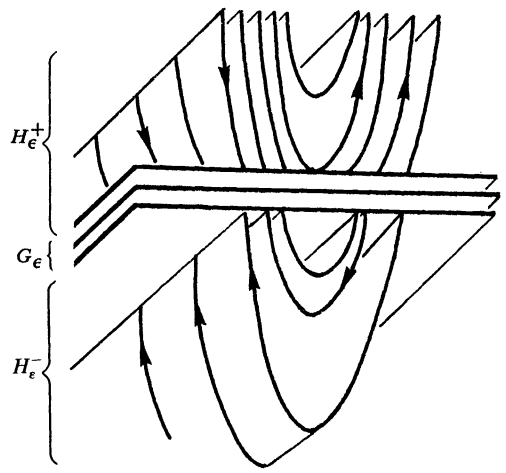


FIG. 5. *Vues sous le macroscope* (VI), *les trajectoires du champ associé à l'équation* $\varepsilon\ddot{x} = x\dot{x}^{[s]}$ *restent proches des paraboles d'équation* $V = x^2/2 + C_0$ *dans les deux halos* H_ε^+ *et* H_ε^-. *Leur ombre est indépendante de* $s \in]0, 2]$.

d'une fonction $(t(x), V(x))$, car $x' = 1$, qui, si $V(x(t_1))$ est limité, vérifie pour tout s, les presque-équations (*)

$$t(x) \simeq t_0 \quad \text{et} \quad V(x) \simeq x^2/2 + V_0$$

où $t_0 = {}^0t_1$ et la constante V_0 est telle que la parabole $V = x^2/2 + V_0$ passe par le point de coordonnées ${}^0(x(t_1), V(x(t_1)))$.

En revenant à l'échelle initiale, tout saut croissant d'une équation (IV) vérifie donc, pour $t \in [t_1, t_2]$, la presque-équation

ou

$$\begin{aligned} &(\varepsilon/(2-s))\dot{x}(t)^{[2-s]} \simeq x^2(t)/2 + V_0, && \text{si } s \neq 2, \\ &\varepsilon \log \dot{x}(t) \simeq x^2(t)/2 + V_0, && \text{si } s = 2, \end{aligned}$$

appelée *équation des sauts*, où V_0 est une constante qui dépend du saut considéré.

On peut également préciser les origines et extrémités des sauts. Pour cela, il convient de s'assurer que les courbes qui sont les ombres des trajectoires dans G (Fig. 4) et celles qui sont les ombres des trajectoires hors de G vues sous le macroscope (Fig. 5), se recollent bien les unes aux autres de la manière qu'on imagine. Plus précisément, il s'agit de vérifier qu'une trajectoire qui atteint des vitesses grandes en longeant (*) une verticale $t = t_0$ et $x = x_-$ vérifiera l'équation des sauts ci-dessus avec $V_0 = -x_-^2/2$, c'est à dire tel que $V(x_-) = 0$; et de même, que celle qui reprend des vitesses limitées en longeant une verticale $t = t_0$, $x = x_+$, la vérifiera pour $V_0 = -x_+^2/2$, c'est à dire que $V(x_+) = 0$. Ces propriétés de recollement sont, comme nous le verrons au chapitre 3, des conséquences du lemme de l'ombre courte (*).

On en déduit facilement les deux résultats suivants:

1°) Tout saut décroissant d'origine ou d'extrémité finie est également d'extrémité ou d'origine finie respectivement, et ces abscisses x_- et x_+ sont reliées par le relation *sortie-entrée* $x_-^2/2 = x_+^2/2$, c'est à dire

$$x_+ = -x_-.$$

2°) Aucun saut croissant n'est borné. Il est soit d'origine $-\infty$ et d'extrémité $x_- \leqq 0$, soit d'origine $x_+ \geqq 0$ et d'extrémité $+\infty$.

On peut également déduire de l'étude précédente l'épaisseur des sauts. Contrairement à l'équation des sauts ou à la relation sortie-entrée, l'épaisseur, elle, dépend de s. Calculons à titre d'exemple cette épaisseur $\mathscr{E}$ dans le cas $s=1$, pour un saut borné (donc décroissant) d'origine x_- et d'extrémité x_+. Par définition, lorsque t est élément de $\mathscr{E}$, $\dot{x}(t)$ est grande et donc $x(t)$ vérifie l'équation des sauts

$$\varepsilon\dot{x}(t) \simeq \left(x^2(t) - x_-^2\right)/2$$

On en déduit, en vertu de la définition de $\mathscr{E}$, que t est un élément de $\mathscr{E}$ si et seulement si $\varepsilon\dot{x}(t) \not\simeq 0$ (c'est à dire qu'au cours du saut, la vitesse est d'ordre $1/\varepsilon$). Comme

$$t = t_1 + \int_{x(t_1)}^{x(t)} \frac{dx}{\dot{x}(t)} = t_1 + \varepsilon \int_{x(t_1)}^{x(t)} \frac{dx}{\varepsilon\dot{x}(t)},$$

t appartient à $\mathscr{E}$ si et seulement si $t = t_1 + \tau$ avec τ/ε limité. En d'autres termes, $\mathscr{E}$ est la ε – galaxie de t_1.

Dans le cas d'un saut non borné dont l'une des extrémités est finie, un raisonnement semblable conduit, toujours pour $s=1$, à la conclusion que $\mathscr{E}$ est la réunion de la galaxie précédente et de la demi-droite $]-\infty, t_1]$, ou $[t_1, +\infty[$. Notons que l'épaisseur d'un saut borné (respectivement non borné) d'extrémité finie est identique pour tous les sauts d'une même équation. On verra au chapitre suivant que cette propriété est générale et que seul le type de croissance de l'équation détermine l'épaisseur des sauts.

Problèmes aux limites. Lutz et Sari ont indiqué [17] comment tirer profit des méthodes non standard pour l'étude de problèmes aux limites concernant certaines équations (I). Nous nous proposons d'examiner, à propos de l'exemple des équations (IV), ce que peut apporter une bonne connaissance des sauts des solutions pour la résolution d'un problème aux limites.

Considérons le problème aux limites suivant associé à (IV):

$$\text{(VII)} \qquad \begin{aligned} &\varepsilon\ddot{x} = x\dot{x}^{[s]}, \quad 1 \leqq s \leqq 2, \\ &x(-1) = a, \quad x(+1) = b. \end{aligned}$$

Traduit en termes de trajectoires de l'espace des phases, ce problème consiste à trouver une trajectoire du champ associé qui joint un point de la verticale d'équation $t=-1$, $x=a$, à un point de la verticale d'équation $t=+1$, $x=b$. La détermination du comportement selon les valeurs de a et b d'une solution de (VII), si elle existe, découle facilement des quelques observations suivantes.

1°) Les solutions de (IV) sont toutes, à l'exception des solutions constantes $x=x_0$, soit croissantes, soit décroissantes (i.e. les deux demi-espaces $v>0$ et $v<0$ sont invariants pour le champ (V) associé à (IV)).

2°) Aucune variation appréciable de x n'est possible à vitesse limitée, pour x limité.

3°) A l'exception des solutions dont les trajectoires associées ont pour ombre, sous le macroscope, une parabole d'équation $V = x^2/2$ ($V_0 = 0$), aucune solution de (IV) ne peut présenter plus d'un saut, les premières présentant, quant à elles, au plus deux sauts.

La Fig. 6 indique, selon les valeurs de a et de b ces comportements nécessaires. En procèdant comme dans [16], on établit l'existence de solutions du problème aux limites

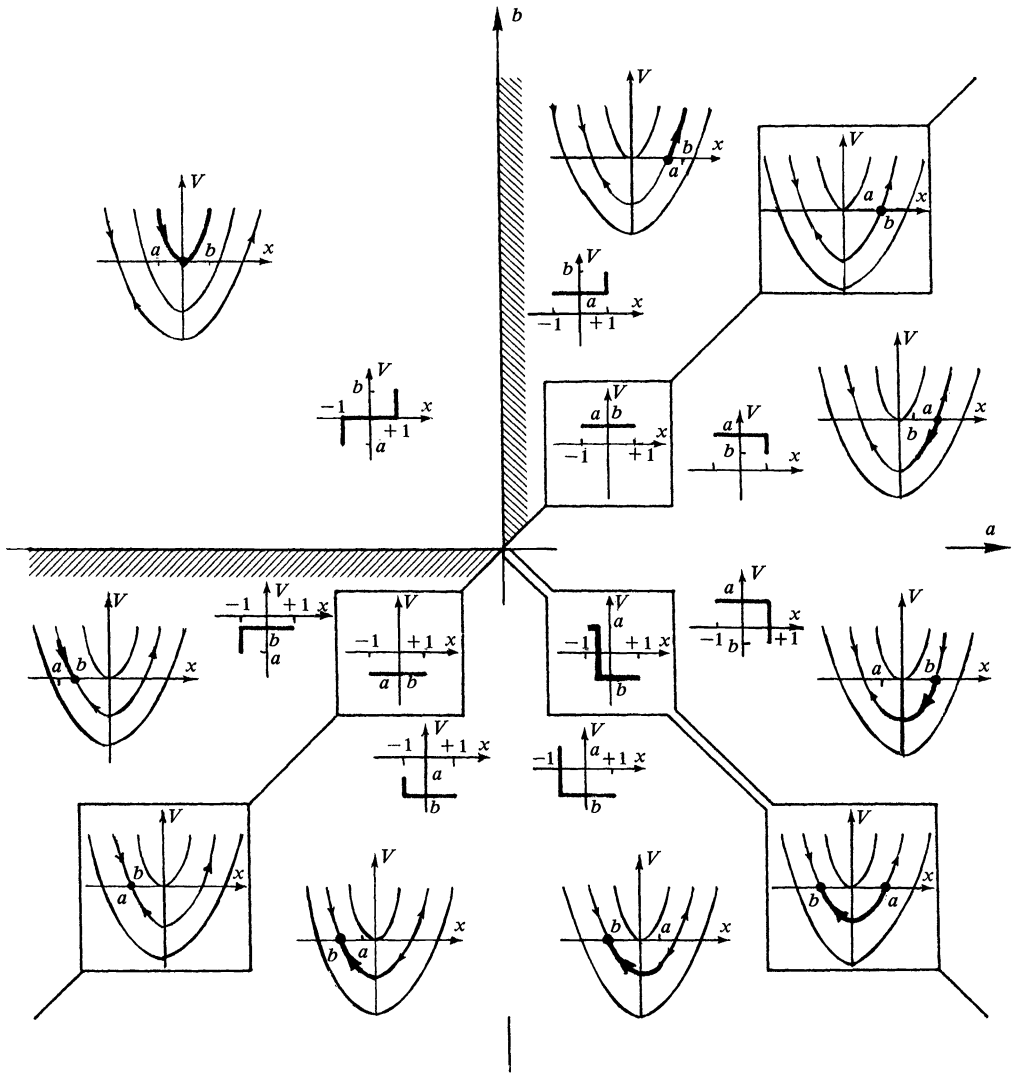


FIG. 6. *Solutions du problème aux limites* $\varepsilon\ddot{x} = x\dot{x}^{[s]}$, $x(-1) = a$, $s(+1) = b$, *selon les valeurs de a et b. Au centre, les graphes* $(t, x(t))$ *de ces solutions; à l'extérieur, les trajectoires associées* $(x, V(x))$, *vues sous le macroscope* (VI). *Du point de vu des sauts, on peut grouper ces solutions en cinq régions*:

Région 1: $a \simeq b$. *Pas de saut.*

Région 2: ${}^0b \leqq 0$, $b < -a$, $b \not\simeq \pm a$. *Un saut à l'instant* -1, *d'origine a et d'extrémité b, et d'équation* $V(t) \simeq x^2(t)/2 - b^2/2$.

Région 3: ${}^0a \geqq 0$, $b > -a$, $b \not\simeq \pm a$. *Un saut à l'instant* $+1$, *d'origine a et d'extrémité b, et d'équation* $V(t) \simeq x^2(t)/2 - a^2/2$.

Région 4: $a \simeq b$, $a > 0$. *Un saut à un instant* $t_0 \in [-1, +1]$ *d'origine a et d'extrémité b, et d'équation* $V(t) \simeq x^2(t)/2 - a^2/2$. *Si* $a = b$, $t_0 = 0$.

Région 5: ${}^0a > 0$, ${}^0b < 0$. *Un saut à l'instant* $t = -1$ *d'origine a et d'extrémité* 0 *et un saut à l'instant* $t = +1$ *d'origine* 0 *et d'extrémité b, tout deux d'équation* $V(t) \simeq x^2(t)/2$.

(VII) présentant chacun de ces comportements, comme un conséquence du théorème de la valeur intermédiaire (shooting) et du lemme de l'ombre courte. Les résultats obtenus ainsi sont connus lorsque $s = 1$. Ce qui est remarquable est qu'ils restent valables pour les autres valeurs de $s \in]0, 2]$.

Ajoutons un commentaire concernant les solutions du problème (VII) présentant une "couche libre" (région 4 de la Fig. 6). On ne trouve pas dans les études antérieures de solutions de (VII) présentant un saut "intérieur" à l'instant $t_0 \neq 0$. Toutefois, pour des raisons de continuité, on peut trouver de telles solutions en choisissant des conditions limités a et b, non toutes deux standard, mais telles que $a \simeq -b$. En termes classiques, l'équivalent de telles conditions aux limites serait obtenu en considérant a et b comme des fonctions $a(\varepsilon)$ et $b(\varepsilon)$ de ε. Il serait d'ailleurs intéressant de calculer une estimation de l'instant de saut t_0 en fonctions de a et de b (et de ε).

3. Equation des sauts. Les solutions d'une équation de type (I) se composent, en général, à la fois de portions parcourues à vitesse limitée, et de portions parcourues à vitesse grande. Leur étude comporte donc deux parties distinctes: l'une concernant les portions lentes, l'autre concernant les sauts. La première est généralement bien connue, car on dispose dans ce cas de l'équation réduite:

$$f(t,x,\dot{x})=0.$$

En effet, si la vitesse est limitée, les solutions sont "bien approchées" (à des sauts de vitesse près) par les solutions de cette équation du premier ordre, comme on peut le voir en examinant les trajectoires associées dans l'espace des phases. En d'autres termes, si la vitesse est limitée, "on peut négliger ε".

Par contre, il n'en est plus ainsi lorsque la vitesse est grande, et le problème se pose de trouver l'analogue, dans ce cas, de l'équation réduite, c'est à dire de trouver une équation, si possible facile à calculer à partir de (I), dont les solutions seraient de "bonnes approximations" des sauts. L'existence d'une telle équation fait l'objet du premier théorème, que nous nous bornons à énoncer dans le cas des sauts croissants.

THEOREME 1 (équation des sauts). *Soit $f(t,x,v)$ une fonction ayant pour type de croissance en v la fonction $F:[v_0,+\infty[\to\mathbb{R}^+$ vérifiant la condition* (III), *et pour mantisse la fonction $a(t,x)$. Soit $x(t)$ une solution croissante limitée de l'équation $\varepsilon\ddot{x}=f(t,x,\dot{x})$ présentant un saut sur l'intervalle $[t_1,t_2]$. Il existe un changement d'échelle de vitesse*

$$v=h(V/\varepsilon)$$

tel que, pour tout $t\in[t_1,t_2]$, $x(t)$ satisfait à

(i)

$$(V(x(t))=:)\,\varepsilon h^{-1}(\dot{x}(t))\simeq V_1+\int_{x(t_1)}^{x(t)}a(t_1,s)\,ds$$

où $V_1=\varepsilon h^{-1}(\dot{x}(t_1))$.

La fonction h est le difféomorphisme de classe S_0 de $[0,+\infty[$ sur $[v_0,+\infty[$ défini par

(ii) $h'=F(h)/h$ *et* $h(0)=v_0$.

Preuve. Il convient tout d'abord de s'assurer que la fonction h définie par (ii) satisfait bien aux propriétés indiquées: h est de classe S_0, car F est de classe S_0 et que $h(0)=v_0$ est limité; par hypothèse F est strictement positive, donc h est strictement croissante, et h est non bornée, car l'ordonnée d'une asymptote horizontale serait un zéro de F. Elle est prolongeable sur tout $[0,+\infty[$ en vertu du lemme suivant:

LEMME. *Soit $F:[v_0,+\infty[\to\mathbb{R}_*^+$ une fonction de classe C^1 et soit h la solution de l'équation différentielle $h'=F(h)/h$ telle que $h(0)=v_0$. Alors h est prolongeable jusqu'à $+\infty$ si et seulement si F vérifie la condition* (III).

Preuve du lemme. Notons $[0,y_0[$ l'intervalle maximal d'existence de la solution h considérée. De l'égalité $1=hh'/F(h)$, on déduit que

$$y=\int_0^y \frac{h(s)h'(s)}{F(h(s))}ds$$

pour tout $y\in[0,y_0[$, ou encore, en posant $u=h(s)$,

$$y=\int_{v_0}^{h(y)} \frac{u}{F(u)}du.$$

Or y_0 est fini si et seulement si $h(y)$ tend vers l'infini quant y tend vers y_0, d'où le lemme.

La fonction h a donc bien les propriétés annoncées. Il reste à montrer que les sauts vérifient la presque-équation(i). Pour cela, on considère le champ (II) associé à l'équation dans l'espace des phases. Le macroscope $h=v(V/\varepsilon)$ transforme ce champ en

$$\begin{aligned}
&\dot{t}=1,\\
&\dot{x}=h(V/\varepsilon),\\
&\dot{V}=f(t,x,h(V/\varepsilon))/h'(V/\varepsilon).
\end{aligned}$$

Soit $x(t)$ un saut sur l'intervalle $[t_1,t_2]$ et $(t,x(t),V(t))$ la trajectoire correspondante, vue sous le macroscope. Au cours du saut, t reste équivalent à t_1 et, comme $\dot{x}=v=h(V/\varepsilon)$ ne peut s'annuler, la trajectoire est le graphe d'une fonction $x\mapsto(t(x),V(x))$ vérifiant $t(x)\simeq t_1$ et

$$V(x)\simeq V(x(t_1))+\int_{x(t_1)}^{x} \frac{f(t_1,s,h)(V(s)/\varepsilon)}{h(V(s)/\varepsilon)h'(V(s)/\varepsilon)}ds.$$

Comme, au cours du saut, la vitesse $v=h(V/\varepsilon)$ est grande, comme h vérifie l'équation (ii), et comme, par hypothèse sur f on a

$$f(t,x,v)/F(v)\simeq a(t,x)$$

pour tout v grand, on en déduit que

$$V(x)\simeq V(x(t_1))+\int_{x(t_1)}^{x} a(t_1,s)\,ds$$

d'où (i). CQFD.

Commentaires. 1) Ce théorème assure l'existence d'une fonction h qui ne dépend que du type de croissance de f par rapport à $\dot{x}$: elle porte le nom *d'ajustement*. La connaissance de h permet, en effet, de "règler" le macroscope afin de "ramener dans son champ de vision" les sauts qui initialement, dans l'espace des phases, sont trop éloignés pour pouvoir être observés. On a, par exemple, si $F(v)=1$ (cas quasi linéaire) et si on choisit $v_0=1$, $h(y)=(2y+1)^{1/2}$. Si $F(v)=v$ (cas semi linéaire) et si on choisit $v_0=1$, $h(y)=y+1$. Si $F(v)=v^2$ (cas quadratique) et si on choisit $v_0=1$, on a $h(y)=\exp y$. Enfin, si $F(v)=v^2\log v$ et si on choisit $v_0=e$, on a $h(y)=\exp(\exp y)$. On voit sur ces exemples que le choix de v_0 est largement arbitraire, et donc aussi celui de h. De plus, pour les deux premiers examples, il semblerait à priori plus naturel de choisir le macroscope $v=h(V/\varepsilon)$ tel que h soit la solution de "l'équation d'ajustment" $h'=F(h)/h$ vérifiant $h(0)=0$, c'est à dire les macoscope $v=(V/\varepsilon)^{1/2}$ et $v=V/\varepsilon$ respectivement. Cependant, un tel choix de $h(0)$ conduirait à un "ajustement" identiquement nul

pour le troisième exemple, de la même façon que le choix $h(0)=1$ conduirait à un ajustement constant et égal à 1 pour le dernier exemple, ce qui ne peut évidemment pas convenir.

2) Géométriquement, ce théorème indique l'existence d'un changement d'échelle permettant de "voir" simultanément dans le plan $(t=t_0, x, V)$ les ombres de tous les sauts à l'instant t_0 de l'équation considérée. On désigne ces plans du nom de *plans d'observabilité des sauts* [3]. On notera que dans un tel plan d'observabilité les sauts croissants ($V>0$) et les sauts décroissants ($V<0$) n'obéissent pas nécessairement aux même règles. Cependant, si pour tous t et x (limités), la fonction $f(t,x,v)$ est impaire par rapport à v (pour $|v|\geqq v_0$, v_0 standard), comme dans l'exemple de chapitre 2, les ombres des sauts dans leur plan d"observabilité sont des courbes qui se déduisent l'une de l'autre par translation verticale. Si pour tout t et x la fonction $f(t,x,v)$ est paire par rapport à v (pour $|v|\geqq v_0$, v_0 standard), les ombres des sauts dans leur plan d'observabilité pour $V>0$ se déduisent de celles pour $V<0$ par symétrie par rapport à l'axe $V=0$ (Fig. 7). On notera également que, lorsque, l'équation considérée est autonôme, les ombres des sauts dans le plan d'observabilité $t=t_0$ sont indépendantes de t_0. Enfin, les ombres des sauts sont des graphes de fonctions V de x croissantes ou décroissantes selon que $a(t_0,x)$ est positive ou négative, et les extréma de ces fonctions sont situés sur les droites $x=x_0$, où x_0 est un zéro de $a(t_0,x)$. Il en résulte que, si pour une solution $x(t)$, $\dot{x}(t)$ est extrémal au cours d'un saut, alors $x(t)$ est à cet instant t_1 dans le halo d'un zéro de $a({}^0t_1,x)$: on a ${}^0(a(t_1,x(t_1)))=0$.

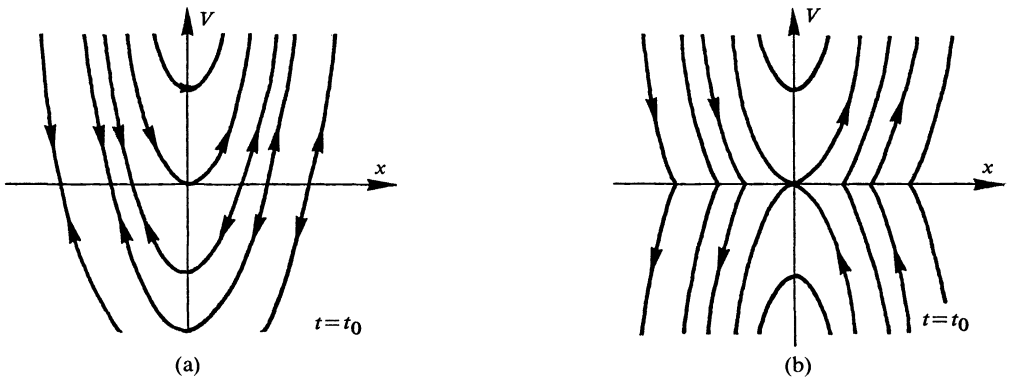


FIG. 7. a) *Trajectoires de l'équation* $\varepsilon\ddot{x}=x\dot{x}^{[2]}$ *dans le plan d'observabilité* $t=t_0$. b) *Trajectoires de l'équation* $\varepsilon\ddot{x}=x\dot{x}^2$ *dans le plan d'observabilité* $t=t_0$.

3) On connait depuis fort longtemps un analogue du plan d'observabilité pour l'étude globale des solutions (et donc en particulier celle des sauts) dans le cas particulier des équations quasi linéaires et autonômes, qu'on appelle encore équations de Lienard:

$$\varepsilon\ddot{x}=a(x)\dot{x}+b(x).$$

On étudie en effet d'ordinaire les solutions de ces équations non dans leur plan de phases $(x, v=\dot{x})$ mais dans leur "plan de Liénard" (x,u) obtenu en posant $u=\varepsilon\dot{x}-A(x)$ où $A(x)=\int_0^x a(s)ds$. A cette échelle, le champ associé est le champ lent-rapide

$$\varepsilon\dot{x}=u+A(x), \qquad \dot{u}=b(x)$$

dont les trajectoires ont des ombres constituées de segments de demi-droites horizontales et de portions de la courbe $u=-{}^0A(x)$. On vérifie facilement que les sauts de

l'équation initiale correspondent précisément aux portions presque-horizontales des trajectories de ce champ. En fait le plan de Liénard (x,u), et le plan d'observabilité $(x, V=\varepsilon v)$ jouent des rôles tout à fait semblables dans l'étude des sauts (ils se correspondent par le difféomorphisme presque-standard $u=V-A(x)$). Cependant, le premier n'existe que pour des équations quasi linéaires, alors que le second peut être utilisé pour toute équation (I) qui vérifie les hypothèses du Théorème 1.

4) L'équation des sauts (i) renseigne également sur "l'ordre de grandeur de la vitesse" au cours des sauts: on constate qu'il existe deux types de sauts: ceux qui, dans leur plan d'observabilité sont d'ordonnée limitée en un point et qui le restent alors en tout point (cas où V_1 est limité), et ceux qui sont d'ordonnée V grande quelque soit x (cas où V_1 est grand). On a donc le résultat suivant:

COROLLAIRE. *Soit f vérifiant les hypothèses du théorème* 1. *Si $x(t)$ est une solution de l'équation $\varepsilon\ddot{x}=f(t,x,\dot{x})$ ayant, à un instant t au moins, à la fois une position et une vitesse limitées, alors la vitesse au cours du saut de x reste nécessairement contenue dans la galaxie des réels pouvant s'écrire $h(\theta/\varepsilon)$, avec θ limité.*

On notera que cette galaxie, qui indique l'ordre de grandeur des vitesses au cours des sauts, ne dépend que du type de croissance de f, et non de la solution ou du saut particulier considéré.

Le théorème 1 indique le comportement des solutions *pendant* les sauts. Nous allons préciser maintenant leur comportement aux "extrémités des sauts". Pour cela, nous supposerons que les sauts considérés ont au moins une extrémité finie et donc, en vertu du corollaire précédent, l'ordonnée $V=h(\dot{x}(t)/\varepsilon)$ reste limitée sur la trajectoire correspondante. Nous désignerons par $A(x)$ la quantité

$$\text{(VIII)} \qquad A(x)={}^0V_1+\int_{0_{x_1}}^{x} {}^0a({}^0t_1,s)\,ds, \quad \text{où } x_1=x(t_1) \text{ et } V_1=h^{-1}(\dot{x}(t_1)).$$

Pour chaque couple (x_1, V_1), A est une fonction standard, continue, définie pour tout x. Le théorème 1 affirme qu'au cours des sauts, les ombres des solutions, dans leur plan d'observabilité, sont contenues dans l'une des courbes d'équation $V=A(x)$. Dans le théorème 2 ci-dessous, nous nous proposons de montrer à présent que si un saut a pour origine x_- (ou pour extrémité x_+), son ombre est contenue dans celle de ces courbes qui passe par le point $(x,V)=(x_-,0)$ (ou $(x,V)=(x_+,0)$). Réciproquement, nous établissons que, si son ombre passe par un tel point, sous de bonnes hypothèses, le saut a x_- pour origine (ou x_+ pour extrémité).

Pour simplifier, le théorème 2 est énoncé dans le cas d'un saut croissant. L'énoncé pour un saut décroissant s'en déduit facilement.

THEOREME 2. *Soit $f(t,x,v)$ une fonction ayant pour type de croissance par rapport à v une fonction F vérifiant la condition* (III), *et $a(t,x)$ pour mantisse. Soit $x(t)$ une solution maximale de l'équation $\varepsilon\ddot{x}=f(t,x,\dot{x})$ présentant un saut croissant sur l'intervalle $[t_1,t_2]$.*

1) *Si ce saut a x_- pour origine, alors $A(x_-)=0$; s'il a x_+ pour extrémité, alors $A(x_+)=0$.*

2) *Réciproquement, s'il existe $x_-<{}^0x_1$ tel que $A(x_-)=0$, si de plus $A(x)\neq 0$ pour tout $x\in]x_-, {}^0x_1]$ et si $a(t_1,x_-)\neq 0$, alors x_- est l'origine du saut.*

De même, s'il existe $x_+>{}^0x_1$ tel que $A(x_+)=0$, si de plus $A(x)\neq 0$ pour tout $x\in[{}^0x_1, x_+[$ et si $a(t_1,x_+)\neq 0$, alors x_+ est l'extrémité du saut.

Preuve. Voici la démonstration du théorème relative à l'origine du saut; on en déduit facilement la démonstration correspondante pour l'extrémité.

On utilise le macroscope $v=h(V/\varepsilon)$ introduit au théorème 1. Sous ce macroscope, le champ associé à l'équation $\varepsilon\ddot{x}=f(t,x,\dot{x})$ dans l'espace des phases devient

$$\begin{aligned}\dot{t}&=1,\\ \dot{x}&=h(V/\varepsilon),\\ \dot{V}&=f(t,x,h(V/\varepsilon))/h'(V/\varepsilon).\end{aligned}$$

Désignons par H_ε^+ et H_ε^- les images par le macroscope des halos H^+ et H^-, des points de l'espace des phases d'ordonnée v grande, positive et négative, respectivement.

Dans H_ε^- et H_ε^+, du fait que $h(V/\varepsilon)$ $(=v)$ est non nul, le champ a les mêmes trajectoires que le champ $\mathscr{C}$, obtenu en multipliant ses composantes par $1/h(V/\varepsilon)$, ce qui revient à effectuer un changement (non standard) de paramétrage de ses trajectoires. Or, en vertu de la définition de h et des hypothèses sur f, ce champ $\mathscr{C}$ est proche, en tout point de $H_\varepsilon^- \cup H_\varepsilon^+$, d'un champ standard $\mathscr{C}_0$:

$$\mathscr{C}\ \begin{cases}t'=1/h(V/\varepsilon),\\ x'=1,\\ V'=f(t,x,h(V/\varepsilon))/h(V/\varepsilon)h'(V/\varepsilon),\end{cases}\quad\simeq\quad \mathscr{C}_0\ \begin{cases}t'=0,\\ x'=1,\\ V'={}^0a({}^0t_1,x).\end{cases}$$

Soit $x: I\to\mathbb{R}$ une solution (maximale) qui présente un saut croissant sur $[t_1,t_2]\subset I$, à l'instant $t_0={}^0t_1$, d'origine x_-. Par définition de l'origine, $\dot{x}(t)$ est grand dés que $x(t)\gg x_-$ (et que $x(t)\leqq x(t_2)$). Dans l'espace des phases, la portion de trajectoire correspondant à cette solution est donc contenue dans H^+, pourvu que $x_-\ll x(t)\leqq x(t_2)$. Vu sous le macroscope, cette portion de trajectoire est une trajectoire de $\mathscr{C}$; puisque $x'=1$, elle s'écrit donc $\gamma(x)=(t(x),x,V(x))$ avec $\gamma(x)\in H_\varepsilon^+$ pour $x_-\ll x\leqq x(t_2)$. Soit $\gamma_0(x)$ la trajectoire de $\mathscr{C}_0$ issue du point ${}^0\gamma(x_1)$. En intégrant ${}_0$, on constate que $\gamma_0(x)=(t_0,x,A(x))$. Et comme $\mathscr{C}$ et $\mathscr{C}_0$ sont proches sur H_ε^+, on a évidement $\gamma(x)\simeq\gamma_0(x)$, et en particulier $V(x)\simeq A(x)$, pour tout x tel que $x_-\ll x\leqq x(t_2)$, l'intervalle $[x_-,x(t_2)]$ étant limité. Comme pour ces valeurs de x, $V(x)\geqq 0$, il en sera de même de $A(x)$ et donc, par continuité de A, on aura $a(x_-)\geqq 0$. Nous allons montrer que $A(x_-)$ ne peut être strictement positif. En effet, comme A est une fonction standard et continue, il existerait sinon un voisinage standard de x_- sur lequel $A(x)>0$. Soit $[b,c]$ un tel voisinage de x_- avec $c\leqq x(t_2)$ (Fig. 8a). Comme $x_-\ll c\leqq x(t_2)$, on a, d'après ce qui précède, $\gamma(c)\simeq\gamma_0(c)$. Par ailleurs, comme $A(x)>0$ pour tout $x\in[b,c]$, $\gamma_0(x)$ $(=(t_0,x,A(x)))$ reste contenue dans la galaxie G^+ des points d'ordonnée V non petite et positive, tant que $x\in[b,c]$. Comme G^+ est une galaxie contenue dans H^+ (où $\mathscr{C}\simeq\mathscr{C}_0$) et contenant le halo de tous ses points, on en déduit, en vertu du lemme de l'ombre courte (*), que $\gamma(x)$ est défini pour tout $x\in[b,c]$ et qu'elle reste proche de $\gamma_0(x)$ sur tout cet intervalle; en particulier $V(x)\gg 0$ pour $x\in[b,c]$. On en déduit que pour tout t tel que $x(t)\in[b,c]$, $\dot{x}(t)$ est grand, ce qui contredit le fait que x_- est l'origine du saut.

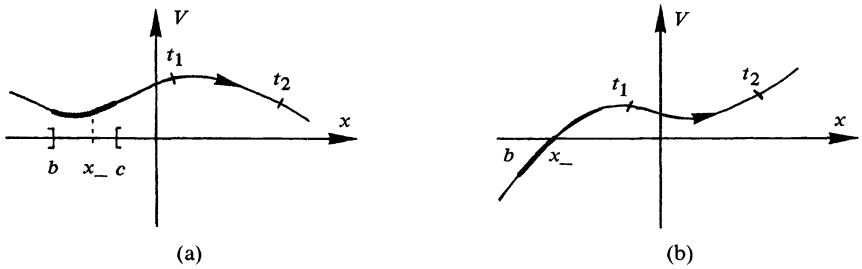


Fig. 8

Montrons la réciproque: Soit $x: I \to \mathbb{R}$ une solution maximale ayant un saut sur l'intervalle $[t_1, t_2]$. Soit $\gamma(x) = (t, x, V(x))$ la trajectoire correspondante du champ $\mathscr{C}$, et $\gamma_0 = (t_0, x, A(x))$ son ombre, trajectoire du champ $\mathscr{C}_0$. Supposons que A s'annule en $x_- < x_1$ $(= x(t_1))$ et qu'elle soit strictement positive pour tout $x \in]x_-, {}^0x_1]$. Comme A est standard et continue, $A(x) \gg 0$ pour $x_- \ll x \leqq {}^0x_1$ et donc aussi $V(x)$. Il en résulte que, pour ces valeurs de x, γ (et γ_0) restent dans H_ε^+ et donc $\dot{x}(t)$ est grande.

D'autre part, comme par hypothèse ${}^0(a(t_1, x_-)) \neq 0$, la fonction standard A a une dérivée non nulle au point standard x_-. Il existe donc un standard $b < x_-$ tel que $A(x) \ll 0$ pour $b \leqq x \ll x_-$. Si, par l'absurde, il existait un voisinage standard $\mathscr{V}_0$ de x_- sur lequel $\dot{x}(t)$ est grande et positive, on aurait, en vertu du théorème 1, $V(x) \simeq A(x)$ pour tout $x \in \mathscr{V}_0$ et donc $V(x) < 0$ pour $x \in]b, x_-[\cap \mathscr{V}_0$ tel que $x \neq x_-$, ce qui est absurde. CQFD

COROLLAIRE (fonction sortie-entrée). *Soit f vérifiant les hypothèses du théorème 2. Si $x(t)$ est une solution de l'équation $\varepsilon\ddot{x} = f(t, x, \dot{x})$ qui présente à l'instant t_0 un saut borné, alors l'origine x_- et l'extrémité x_+ du saut sont liées par la relation:*

$$\int_{x_-}^{x_+} {}^0a(t_0, s)\, ds = 0.$$

Commentaires. 1) La seconde partie du théorème 2 a été établie sous l'hypothèse que le zéro de $A(x)$ considéré n'est pas également un zéro de sa dérivée ${}^0a(t_0, x)$. En termes géométriques, cela signifie que dans le plan d'observabilité le courbe d'équation $V = A(x)$ qui contient l'ombre du saut entre $x(t_1)$ et $x(t_2)$ et qui rencontre l'axe $V = 0$ au point d'abscisse x_- (ou x_+) n'est pas tangente à cet axe en ce point. Ce n'est évidemment pas une condition nécessaire pour que x_- soit l'origine du saut, comme le montre l'exemple du chapitre 2 qui présente un saut d'origine $x = 0$, bien que 0 soit un zéro de $a(t_0, x) \eqsim x$ (Fig. 5). Généralement, lorsqu'une courbe $V = A(x)$ dans un plan d'observabilité est tangente à l'axe $V = 0$ en un point d'abscisse x_-, elle contient à la fois l'ombre de sauts d'origine x_- et l'ombre de sauts pour lesquels x_- n'est pas l'origine, soit que cette origine soit strictement inférieure, soit qu'ils soient sans origine (Fig. 9).

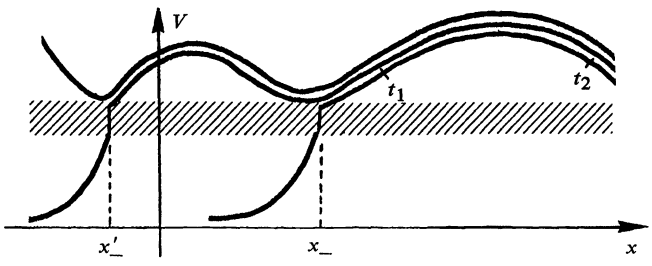


FIG. 9. *Sauts avec "rebond" ou sans "rebond" en x_-.*

2) Lorsque la mantisse $a(t, x)$ est identiquement nulle pour $t = t_0$, ou plus généralement lorsque la fonction $x \mapsto a(t_0, x)$ a des zéros non isolés, les renseignements sur les sauts de l'équation à l'instant t_0 fournis par le théorème 1 sont inutilisables, et le théorème 2 ne s'applique plus. En effet, l'équation possède alors des sauts dont l'ombre, dans leur plan d'observabilité, a pour équation $V = 0$. Il est impossible de déterminer, à cette échelle, l'origine ou l'extrémité de tels sauts; la fonction sortie-entrée n'existe pas

en général (x_+ dépend à la fois de x_- et de la solution considérée), (Fig. 10). Il apparait que le plan d'observabilité n'est pas la "bonne échelle" pour l'étude de ces sauts (voir à ce sujet [7]).

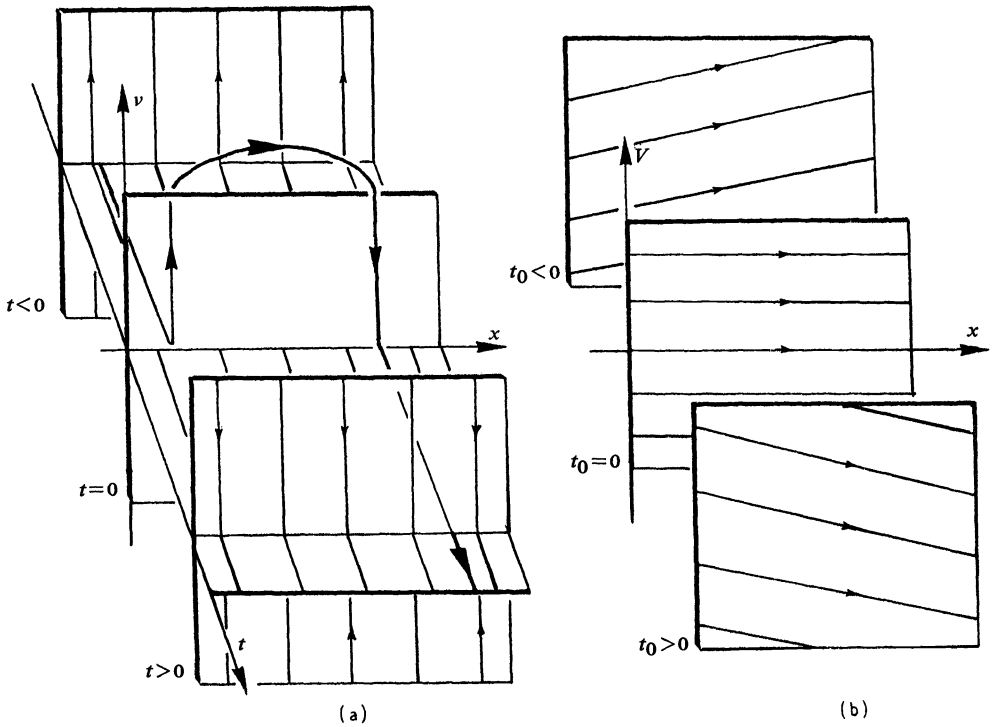


FIG. 10. *Trajectoires du champ associé à l'équation* $\varepsilon\ddot{x} = -t\dot{x}$, *(a) dans l'espace des phases (b) dans des plans d'observabilité (l'équation des sauts est* $V(x) \simeq -t_0 x$*). On observe qu'il ne peut y avoir de saut borné qu'à l'instant* $t_0 = 0$*. L'équation d'un saut borné est alors* $V(x) \simeq 0$*. On ne peut donc pas calculer de fonction sortie-entrée. De fait, en intégrant l'équation, on a* $x(t) = \int_{x(0)}^{x} (C_1 \exp(-t^2/2) + C_2)\, dt$*; on constate que pour tout couple* (x_-, x_+) *de réels standard, il existe une solution présentant un saut à l'instant* $t_0 = 0$ *d'origine* x_- *et d'extrémité* x_+.

4. Epaisseur des sauts. Nous allons voir à présent que les résultats des théorèmes 1 et 2 permettent généralement de calculer l'épaisseur des sauts en fonction de l'ajustement h. Une nouvelle fois, nous nous bornerons au cas des sauts croissants.

Soit h un ajustement (défini au théorème 1) associé à une équation (I), c'est à dire un difféomorphisme de classe S^0 de $[0, +\infty[$ sur $[v_0, +\infty[$ vérifiant l'équation $hh' = F(h)$. Soit $h_\varepsilon : \mathbb{R} \to \mathbb{R}$ l'homéomorphisme croissant défini par

$$h_\varepsilon(y) = \begin{cases} 1/h(1/\varepsilon y) & \text{pour } y > 0, \\ 0 & \text{pour } y = 0, \\ -1/h(-1/\varepsilon y) & \text{pour } y < 0. \end{cases}$$

Soit enfin G_h l'image par h_ε de la galaxie principale $G(*)$. L'ensemble G_h a les propriétés suivantes:

C'est une galaxie: en effet, G_h est par définition une prégalaxie $(*)$; elle est strictement externe car h_ε est continue et prend à la fois des valeurs limitées et des valeurs grandes.

Cette galaxie est convexe (h est continue et non bornée), et est symétrique par rapport à 0 (h est impaire).

Elle est contenue dans le halo de zéro: ceci découle du fait que l'ajustement $h(y)$ est grand pour toute valeur grande de y. En effet, h est de classe S^0 et ${}^0h$ est solution de l'équation standard $h' = {}^0F(h)/h$. Comme h et ${}^0h$ sont proches pour tout y limité, ils sont encore proches pour tout $y \leqq Y$ pour un Y grand (principe de Fehrele). Donc $h(y)$ est grand pour tout y grand inférieur à Y; par monotonie de h, cela reste donc vrai pour les y supérieurs à Y également.

C'est un sous-groupe (externe) de $\mathbb{R}$: En effet, comme G_h est convexe et symétrique par rapport à 0, il suffit de vérifier que pour tout $t>0$, si $t \in G_h$ alors $2t \in G_h$. Notons tout d'abord que h^2 est une fonction convexe, puisque

$$(h^2)'' = (2hh')' = (2F(h))' = 2F'(h)h'$$

et donc $(h^2)''$ est positive car F et h sont croissantes. Comme $t>0$, il existe y limité tel que $t = 1/h(1/\varepsilon y)$. Comme G_h est convexe et contient 0, pour montrer que $2t \in G_h$, il suffit de montre que $2t$ est majoré par un élément de G_h, c'est à dire qu'il existe Y limité tel que

$$(2t =) \quad 2/h(1/\varepsilon y) \leqq 1/h(1/\varepsilon Y)$$

ou encore, tel que

$$4h^2(1/\varepsilon Y) \leqq h^2(1/\varepsilon y).$$

Mais comme h^2 est convexe, on a, pour tout z,

$$h^2(z) \leqq \big(h^2(2z) + h^2(0)\big)/2$$

ou encore, en réappliquant deux fois cette inégalité,

$$h^2(z) \leqq \big(h^2(8z) + 7h^2(0)\big)/8.$$

Comme $h^2(0) = v_0^2$ est limité et que $h^2(8z)$ est grand des que z est grand, on en déduit que pour tout z grand,

$$8h^2(z) \leqq 2h^2(8z).$$

Il suffit donc de poser $Y = 8y$.

Voici quelques exemples de galaxies G_h:

Si $h(y) = y$, G_h est la ε-galaxie de 0. Plus généralement, si $h(y) = y^r$, G_h est la ε^r-galaxie de 0; c'est donc une galaxie linéaire. Par contre, si $h(y) = \exp(y)$, G_h est la ε–micro-galaxie; c'est donc une galaxie non linéaire.

Comme les deux précédents, le théorème 3 sera énoncé dans le cas d'un saut croissant:

THEOREME 3 (épaisseur des sauts). *Soit $f(t,x,v)$ une fonction ayant un type de croissance F par rapport à v vérifiant la condition* (III). *Soient h un ajustement correspondant, et G_h la galaxie associée. Soient $x(t)$ un saut borné de l'équation $\varepsilon\ddot{x} = f(t,x,\dot{x})$ d'épaisseur noté $\mathscr{E}$, et t_1 un élément de $\mathscr{E}$. Supposons que*

(i) *la fonction $A(x)$ définie par* (VIII) *ne s'annule en ancun point strictement compris entre les extrémités du saut,*

(ii) *la fonction $H(V) =: h(V/\varepsilon)/h(1/\varepsilon)$ soit de classe S^1 en $V = 0$.*

Alors $\mathscr{E}=t_1+G_h$ (*= image de* G_h *par la translation* $t\mapsto t+t_1$).

Preuve. Comme $\dot{x}(t)=v$, et qu'au cours d'un saut, $v(t)$ peut s'exprimer comme fonction de $x(t)$, on a

$$t=t_1+\int_{x(t_1)}^{x(t)}\frac{ds}{v(s)}=t_1+\int_{x(t_1)}^{x(t)}\frac{ds}{h(V(s)/\varepsilon)} \quad \text{car } v=h(V/\varepsilon).$$

Posons

$$e(x)=\int_{x_1}^{x}\frac{ds}{h(V(s)/\varepsilon)}, \quad \text{où } x_1=x(t_1).$$

Soient x_- et x_+ les extrémités du saut, et notons

$$E=\{\, x \mid x_- \ll x \ll x_+ \,\} \quad \text{et} \quad \mathscr{E}_0=e(E).$$

On a $\mathscr{E}=t_1+\mathscr{E}_0$, et le théorème affirme que $\mathscr{E}_0=G_h$.

Notons tout d'abord que $\mathscr{E}_0$ et G_h sont des convexes contenant 0, et donc, pour que t soit un élément d'un tel ensemble, il suffit qu'il existe t' dans cet ensemble tel que $t'/t\geqq 1$.

1) Montrons tout d'abord que $\mathscr{E}_0\subset G_h$: Par définition des extrémités x_+ et x_-, $v(x)$ est grand pour tout $x\in E$ donc, en vertu du théorème 1, $V(x)\simeq A(x)$ pour $x\in E$. Du fait de l'hypothèse (i), $A(x)\neq 0$ pour tout $x\in E$ et, en fait, $A(x)\gg 0$ puisque le saut est supposé croissant et que A est standard et continue. Donc $V(x)\gg 0$ pour tout $x\in E$. On en déduit que, pour tout $x\in E$, $1/h(V(x)/\varepsilon)$ est un élément de G_h. A présent, pour tout $x_0\in E$, on a par le théorème de la moyenne, que $e(x_0)=(x_0-x_1)/h(V(x)/\varepsilon)$ pour un $x\in E$. Comme $1/h(V(x)/\varepsilon)\in G_h$, que G_h est un groupe, et que (x_0-x_1) est limité, on a $e(x_0)\in G_h$. D'où l'inclusion recherchée.

2) Montrons que réciproquement $G_h\subset\mathscr{E}_0$: Pour cela, nous allons montrer que pour tout $T\in G_h$, $T>0$, il existe $t\in\mathscr{E}_0$ tel que $t>T$, c'est à dire qu'il existe $x\in E$ tel que $e(x)>T$. On procèderait de manière analogue pour $T<0$.

Envisageons d'abord le cas où G_h est une galaxie linéaire, c'est à dire qu'il existe α tel que $G_h=\alpha\mathbb{G}$, où $\mathbb{G}$ est la galaxie principale. De par la définition de G_h, on peut choisir $\alpha=1/h(1/\varepsilon)$. Il convient donc de prouver, sous cette hypothèse, que pour tout y limité, il existe $x\in E$ tel que $h(1/\varepsilon)\cdot e(x)>y$, ou encore, tel que

$$\int_{x_1}^{x}\frac{ds}{H(V(s))}>y.$$

Or la fonction $H(V)$ est limitée pour V limité, et non petite pour $V\gg 0$. Comme par hypothèse cette fonction dérivable est de classe S^1 en $V=0$, il existe un intervalle standard $[0, V_0]$ sur lequel H reste proche d'une fonction standard H_0, dérivable en $V=0$, et nulle en ce point car $H(0)\simeq 0$. Son inverse $1/H_0$ est donc d'intégrale divergente en $V=0$. Ceci va nous permettre de conclure. En effet, comme H (et donc H_0) est non petite pour $V\gg 0$, $1/H(V(x))\simeq 1/H_0(V(x))$, pour tout $x\in E$. Or l'ensemble des x pour lesquels l'équivalence ci-dessus est vérifiée est un halo qui contient la galaxie E. Par le principe de Fehrele, il contient également des $x\notin E$. Il existe donc $x_0\simeq x_+$ tel que cette équivalence soit satisfaite pour tout $x\in[x_1, x_0]$. Mais comme x_+ est l'extrémité du saut, on a $V(x_+)\simeq 0$ (théorème 2). Donc, pour tout $x\leqq x_0$, l'intégrale $\int_{x_1}^{x}(ds/H(V(s)))$ est équivalente à l'intégrale $\int_{x_1}^{x}(ds/H_0(V(s)))$ qui est grande dès que $x\simeq x_+$. En effet, si $x\simeq x_+$, $V(x)\simeq 0$ (théorème 2). Par continuité, pour tout y limité, il

existe donc $x \ll x_+$ tel que

$$\int_{x_1}^{x} \frac{ds}{H(V(s))} > y$$

ce qui achève la démonstration dans le cas où G_h est linéaire.

Considérons à présent le cas où G_h est une galaxie non linéaire. La non linéarité de G_h peut s'exprimer par la propriété suivante [1]:

"Pour tout $T \in G_h$, il existe $T' \in G_h$ tel que T'/T est grand". Comme la quantité $1/h(V(x)/\varepsilon)$ est une fonction croissante de x pour x assez proche de x_+, et comme elle appartient à G_h tant que $x \ll x_+$ et qu'elle quitte G_h dès que $x \simeq x_+$ (car $V(x) \simeq 0$, d'après le théorème 2), pour tout $T' \in G_h$ il existe une valeur $x_0 \ll x_+$ telle que $1/h(V(x)/\varepsilon) > T'$ pour tout $x \in [x_0, x_+]$. Soit x_0' tel que $x_0 \ll x_0' \ll x_+$. On a

$$e(x_0') = \int_{x_1}^{x_0'} \frac{ds}{h(V(s)/\varepsilon)} > \int_{x_0}^{x_0'} \frac{ds}{h(V(s)/\varepsilon)} > \int_{x_0}^{x_0'} T' ds > T. \qquad \text{CQFD}$$

Remarque. On notera que dans le cas où G_h est non linéaire, l'hypothèse (ii) du théorème n'intervient pas dans la démonstration. Il semble qu'elle soit toujours satisfaite dans ce cas.

Commentaires. 1) Ce théorème montre en particulier que, sous les hypothèses (i) et (ii), tout saut borné d'une équation (I) peut être "étalé" sur la galaxie principale par le changement de temps $\tau \mapsto t = t_1 + 1/h(1/\varepsilon\tau)$. Dans le cas quasi linéaire ($h(y) = y$), par exemple, on retrouve le changement de variable adopté habituellement: $t - t_1 = \varepsilon\tau$. Du théorème, il découle également que les épaisseurs de sauts bornés sont les images par des translations, de sous groupes convexes de $\mathbb{R}$ (si $t \in \mathscr{E}$, $2(t - t_1) + t_1 \in \mathscr{E}$) et qu'il s'agit de galaxies linéaires si et seulement si G_h est linéaire. On comprend à présent pourquoi il n'est pas toujours possible "d'étaler" les sauts par des changements de temps du type $t - t_1 = 1(\varepsilon)\tau$: ce n'est possible que lorsque G_h est linéaire. En particulier pour les équations dont le type de croissance $F(v)$ est une fonction puissance $x \mapsto x^s$, la linéarité de G_h équivaut à la condition $s < 2$ (h croit "nettement moins vite" que la fonction exponentielle). Dans le cas général, il serait utile de diposer d'un critère simple sur le type de croissance F de l'équation qui équivaille à la linéarité de G_h (et donc à la linéarité de l'épaisseur des sauts). Il semble que le critère suivant devrait convenir:

$$G_h \text{ est linéaire} \Leftrightarrow \int_{v_0}^{+\infty} \frac{F(v)}{v^3} dv < \infty,$$

ce qui est, par définition de h, équivalent à

$$G_h \text{ est linéaire} \Leftrightarrow \int_0^{+\infty} \left(\frac{h'(y)}{h(y)} \right)^2 dy < \infty.$$

2) Précisons le sens des hypothèses (i) et (ii) du théorème: La première porte sur les sauts. Elle signifie géométriquement que, dans le plan d'observabilité, l'ombre du saut ne rencontre pas l'axe $V = 0$ en des points d'abscisse strictement comprise entre les extrémités' x_- et x_+, et donc, en quelque sorte, qu'il s'agit d'un saut sans "rebond" Lorsqu'un saut ne vérifie pas cette hypothèse, on déduira facilement du théorème que son épaisseur $\mathscr{E}$ est alors la réunion d'un intervalle petit $I = [t_1, t_2]$ et des deux galaxies

$t_1+G_h^-$ et $t_2+G_h^+$, avec $G_h^{\pm}=G_h\cap\mathbb{R}^{\pm}$, le saut parcourant les divers rebonds pout $t\in[t_1,t_2]$.

La seconde hypothèse porte sur l'équation. Elle signifie que la fonction H (que l'on peut calculer dès que l'on connait F, et donc h) est équivalente sur un voisinage standard de 0 à une fonction standard H_0 dérivable au point $V=0$. Elle est vérifiée, par exemple, lorsque h est l'identité (cas quasi linéaire), car $H(V)=H=V_0(V)$. Elle est aussi vérifiée lorsque h est une exponentielle (cas quadratique) car $H(V)=e^{(V-1)/\varepsilon}$ est infinitésimale pour tout $V\ll 1$, donc $H_0(V)=0$, pour $V\ll 1$. Par contre elle est en défaut lorsque $h(v)=\sqrt{v}$ (cas semi-linéaire) car $H(V)=\sqrt{V}=H_0(V)$ n'est pas dérivable en 0. Plus généralement, si le type de croissance est une fonction puissance $F(v)=v^s$, l'hypothèse (ii) est équivalente à la condition $s\geqq 1$. Il est probable qu'une condition équivalente à (ii) serait:

$$\int_{v_1}^{+\infty}\frac{F(v)}{v^2}dv=+\infty$$

ou encore, en fonction de h:

$$\int_{v_1}^{+\infty}\frac{(h'(y))^2}{h(y)}dy=+\infty.$$

On peut, cependant, se poser la question de savoir quelle est l'épaisseur des sauts lorsque la condition (ii) n'est pas remplie, par exemple pour des équations ayant un type de croissance $F(v)=v^s$ avec $s<1$. L'inclusion $\mathscr{E}\subset t_1+G_h$ reste valable dans tous les cas, mais il arrive que cette inclusion soit stricte. Plus précisément, on peut montrer par un affinement de la preuve précédente, que les conclusions du théorème subsistent pourvu que l'ombre du saut dans son plan d'observabilité soit tangente à l'axe $V=0$ en ses extrémités (voir [3]). Dans le cas contraire, l'épaisseur n'est plus la translaté d'une galaxie sous-groupe, mais c'est une galaxie obtenue à partir d'un intervalle auquel on retire un halo de l'extrémité concernée. Ce halo est le translaté d'un halo sous-groupe; il peut également être exprimé en fonction de h ([3]).

3) Pour simplifier, le théorème précédent a été énoncé dans le cas des sauts ayant à la fois une origine *et* une extrémité finies. En fait, par une adaptation évidente de la preuve qui a été donnée, on établirait un résultat englobant le cas des sauts non bornés, ayant une origine *ou* une extrémité finie. En effet, s'il s'agit par exemple d'un saut d'extrémité finie, alors, sous les hypothèses du théorème, on a montré que

$$\{t\in\mathscr{E},\ t\geqq t_2\}=t_2+G_h^+\quad\text{où } G_h^+=G_h\cap\mathbb{R}^+.$$

De même que, s'il s'agit d'un saut d'origine finie, alors

$$\{t\in\mathscr{E},\ t\leqq t_1\}=t_1+G_h^-\quad\text{où } G_h^-=G_h\cap\mathbb{R}^-.$$

Les épaisseurs de sauts non bornés sont donc la réunion d'un segment (interne) et de l'image par une translation de la partie positive ou de la partie négative de G_h.

Annexe non standard. Des adjectifs tels que petit, grand, appréciable, n'étaient pas d'usage courant en mathématiques jusqu'à l'apparition récente des mathématiques non standard. Pourtant le non mathématicien, physicien ou chimiste par exemple, qui propose à notre sagacité des équations du type (I) n'hésite pas à parler de "petit" paramètre pour désigner ε. Pour lui, en effet, cette petitesse ne fait pas mystère: il s'agit

par exemple d'un paramètre de l'ordre de 10^{-1} ou 10^{-2} alors que les autres paramètres intervenant dans l'équation sont de l'ordre de 10 ou 100. A défaut d'un formalisme adapté, on traduisait généralement la "petitesse" de ε par l'introduction d'une variable ε tendant vers 0 ce qui, en rendant variable une quantité qui ne l'était pas à priori, ne contribuait pas à simplifier le problème. C'est à Reeb [20] que l'on doit l'idée d'introduire l'Analyse Non Standard dans ces questions de perturbations singulières. Ce fut le point de départ de nombre d'études d'équations de type (I), ou plus généralement, de champs lents-rapides. Aux principes originaux de l'analyse non standard, tels qu'on pourra les trouver exposés dans [21] ou [18], se sont ainsi ajoutés petit à petit divers objets ou qualificatifs liés au contexte particulier des équations différentielles, et certains raisonnements omniprésents qui portent aujourd'hui des noms particuliers. C'est l'abondance de ce vocabulaire inhabituel qui me pousse à proposer l'index ci-dessous. Le lecteur n'aura pas manqué de remarquer qu'une part importante des termes utilisés peuvent être interprétés dans un premier temps dans un sens intuitif (celui du physicien, par exemple). C'est un des mérites de l'analyse non standard que de rendre possible cet usage rigoureux de notions jusqu'ici heuristiques. Pour une véritable initiation aux méthodes non standard, nous renvoyons le lecteur à [16] ou [6].

Nous utilisons couramment les symboles suivants:

$\simeq$, $\not\simeq$, $\gg$, $\ll$: voir "équivalent".
${}^0(\,)$: voir "standard".
${}^S(\,)$: voir "standardisé".
$\mathrm{hal}(a)$, $\varepsilon-\mathrm{hal}(a)$: voir "halo".
$\mathbb{G}$, $\varepsilon-\mathrm{gal}(a)$: voir "galaxie".

ajustement: voir théorème 1 et ses commentaires.

appréciable: Se dit d'un réel qui n'est ni petit ni grand (voir ces mots). Un vecteur de $\mathbb{R}^n$ est dit appréciable si ses coordonnées le sont.

Cauchy (principe de): voir "principes de permanence".

changement d'échelle: Difféomorphisme dont la dérivée est grande ou petite. Les plus simples sont les *loupes*, c'est-à-dire les homotéthies (ou affinités) de rapport grand, tels que $x \mapsto x/\varepsilon$. On distingue parmi les changements d'échelle les *microscopes*, qui "grossissent", c'est à dire par lesquels l'image d'une région petite peut être une région appréciable, et les *macroscopes*, qui "rappetissent", c'est à dire par lesquels l'image d'une région grande peut être une région limitée. Ainsi le difféomorphisms $x \mapsto x^{[\varepsilon]} = |x|^{\varepsilon-1}x$ est un macroscope pour $|x| \gg 1$ et un microscope pour $|x| \ll 1$.

classe S^0, S^1 (fonction de): Lorsqu'une fonction interne est non standard, elle peut ou non être proche d'une fonction standard. Si elle est dérivable, la question se pose également pour sa dérivée, etc... Ainsi la fonction $x \mapsto \varepsilon x$ est proche de la fonction nulle, mais la fonction $x \mapsto e^{x/\varepsilon}$ n'est proche d'aucune fonction standard si $x > 0$. Une fonction interne est dite *S-continue* au point x si, pour tout y,

$$y \simeq x \Rightarrow f(y) \simeq f(x).$$

Une fonction interne est de classe S^0 sur une partie E si elle prend des valeurs presque-standard (voir ce mot) aux points presque-standard de E et si elle est S-continue en ces points. Une fonction interne est de classe S^1 sur une partie E si elle est de classe S^0 sur E et si pour tout a presque standard de E, et pour tout x et y équivalents (voir ce mot) à a, le rapport

$$\frac{f(x)-f(y)}{x-y}$$

est presque standard. On montre [6] que toute fonction de classe S^0 est proche d'une unique fonction standard continue, et toute fonction de classe S^1 est proche d'une unique fonction standard de classe C^1. On définit de manière analogue les fonctions de classe S^n [6].

épaisseur (d'un saut): voir chapitre 1.

équivalent: Deux nombres réels sont dits équivalents s'ils diffèrent l'un de l'autre par un infinitésimal (voir ce mot). Ainsi, si ε est petit, 1 et $1+\varepsilon$ sont équivalents. On note $\simeq$ pour "équivalent" et $\not\simeq$ pour "non équivalent" ($2 \not\simeq 3$). La notation $a \ll b$ signifie $a<b$ et $a \not\simeq b$. On dit que a est *nettement inférieur à* b. De même pour $a \gg b$. Deux fonctions f et g sont dites équivalentes sur E si pour tout $x \in E$, $f(x)$ est équivalent à $g(x)$. Si f est interne, l'ensemble des x pour lesquels $f(x) \simeq g(x)$ est un préhalo (voir "halo"). [infinitely close]

externe: Le mathématicien non standard travaille avec deux sortes d'ensembles: les premiers sont les ensembles usuels de la théorie des ensembles de Zermelo–Fraenkel (par exemple): on les appelle *internes*. Ce sont par exemple $\mathbb{N}$, $\mathbb{R}$, ou $[-\omega, +\alpha]$. Les seconds sont ceux qu'on définit en faisant usage de l'adjectif "standard" (ou de ses dérivés "petit", "grand", "limité", $\simeq$, $\ll$) et qu'on appelle *externes*, ou *strictement externes* quand il est établi que le fait de les supposer internes conduirait à une contradiction. Ainsi l'ensemble des réels standard, ou l'ensemble des infinitésimaux sont des ensembles strictement externes (ils sont majorés et n'ont pas de borne supérieure). Une fonction est dite interne si son graphe est un ensemble interne. Ainsi $x \mapsto \varepsilon x$ est interne, mais la fonction qui à x associe la valeur 0 si x est petit, et la valeur 1 sinon est externe.

Extrémité (d'un saut): voir chapitre 1.

Fehrele (principe de): voir "principes de permanence".

galaxie: On appelle *galaxie principale*, et on note $\mathbb{G}$, l'ensemble (strictement externe) des réels limités (voir ce mot). Une *prégalaxie* est l'image réciproque de $\mathbb{G}$ par une fonction f interne. Elle peut être interne (par exemple si $f(x)=1$), ou strictement externe (par exemple si $f(x)=x/\varepsilon$). Dans ce dernier cas, la prégalaxie est appelée une galaxie [2], [8]. Si a est un réel standard, on appelle *ε-galaxie de a*, et on note $\varepsilon-\mathrm{gal}(a)$, l'image réciproque de $\mathbb{G}$ par $f(x)=(x-a)/\varepsilon$, et *ε-microgalaxie de a*, l'image réciproque de $\mathbb{G}$ par $f(x)=1/(\varepsilon \log|x-a|)$. Le complémentaire, dans un interne, d'une galaxie, est un halo (voir également ce mot).

grand: Un réel est dit grand s'il est plus grand que tout entier standard. L'existence d'entiers (et donc de réels) grands est une conséquence du principe d'idéalisation (voir ce mot). [infinite/unlimited]

halo: On appelle halo de 0, noté hal(0), l'ensemble strictement externe des infinitésimaux (voir ce mot). Un *préhalo* est l'image réciproque de hal(0) par une fonction interne f. Il peut être interne (si par exemple f est constante), ou strictement externe (si par exemple $f(x)=x/\varepsilon$). Dans ce dernier cas, le préhalo est appelé un halo [2], [8]. Si a est un réel, on appelle $\varepsilon-$halo de a, et on note $\varepsilon-\mathrm{hal}(a)$, l'image réciproque de hal(0) par $f(x)=(x-a)/\varepsilon$, et ε-microhalo de a, la préimage de hal(0) par $f(x)=1/(\varepsilon \log|x-a|)$. Si A est une partie d'un espace métrique, le halo de A, noté hal(A), est l'ensemble des points à distance petite de A. [monad]

idéalisation: C'est l'un des trois principes fondamentaux de l'analyse non standard I.S.T. [18], [6]. Il s'énonce ainsi: Pour toute formule interne B, contenant au moins deux variables libres x et y, on a:

$$(\forall z,\ z \text{ standard et fini} \Rightarrow \exists x, \forall y \in z,\ B(x,y)) \Leftrightarrow (\exists x_0, \forall y,\ y \text{ standard} \Rightarrow B(x_0,y))$$

On déduit de ce principe l'existence d'entiers grands, et donc de réels petits non nuls.

infinitésimal: Synonyme de "réel petit"' (voir ce mot).

interne: voir "externe".

lemme de l'ombre courte: Soient X et X_0 deux champs de vecteurs localement lipschitziens, définis sur un ouvert de $\mathbb{R}^n$, le premier interne, le second standard. Soit H le (pré)halo où $X \simeq X_0$. Soit I un intervalle compact standard, t_0 un point standard de I, et u_0 une trajectoire de X_0, définie sur I. Soit u une trajectoire de X telle que $u(t_0) \simeq u_0(t_0)$. S'il existe une galaxie G limitée, contenant $u_0(I)$, contenue dans H, et si G contient le halo de tous ses points, alors $u(t)$ est prolongeable à tout I, et on a, pour tout $t \in I$, $u(t) \simeq u_0(t)$ [3].

lemme de Robinson: voir "principes de permanence".

limité: Un réel est dit limité s'il n'est pas grand. Tout réel limité x est équivalent à un unique réel standard, qu'on appelle sa partie standard, et qu'on note ${}^0x$. [finite/limited]

linéaire (galaxie ou halo): Une galaxie (voir ce mot) est dite linéaire si elle est l'image réciproque de $\mathbb{G}$ par une application affine. Ainsi les ε-galaxies d'un point sont linéaires, mais les ε-microgalaxies ne le sont pas. On définit de même les halos linéaires [2], [8].

longer: Une courbe $t \mapsto x(t)$ dans $\mathbb{R}^2$ ou $\mathbb{R}^3$ longe une courbe ou une surface si son image reste à distance petite de cette courbe ou surface.

loupe: voir "changement d'échelle".

macroscope: voir "changement d'échelle".

microscope: voir "changement d'échelle".

ombre courte: voir "lemme de 1".

ombre: L'ombre d'une partie interne A de $\mathbb{R}^n$, notée $[A]$, est le standardisé (voir ce mot) du halo de A. Si la partie A est limitée, son ombre est l'unique ensemble standard compact proche (voir ce mot) de A. Si une fonction est de classe S^0, l'ombre de son graphe est le graphe d'une fonction standard continue. [standard part]

partie standard: voir "limité" et "standard".

petit: Un réel est dit petit si sa valeur absolue est plus petite que tout réel standard strictement positif. [infinitesimal]

plan de Liénard: voir [3] et chapitre 3, commentaire du théorème 1.

plan d'observabilité: voir chapitre 3, commentaire du théorème 1.

prégalaxie: voir "galaxie".

préhalo: voir "halo".

presque-: Dans des expressions telles que "presque-verticale", "presque-horizontale", cet adverbe prend le sens de "à un infinitésimal près". Une presque équation est une relation $R(x)$ obtenue en exprimant que deux termes internes $f(x)$ et $g(x)$ sont équivalents: $f(x) \simeq g(x)$.

principes de permanence: Le plus connu est le *lemme de Robinson* [21]: "Soit (u_n) une suite interne de réels. Si $u_n \simeq 0$ pour tout n limité, alors il existe un N grand tel que $u_n \simeq 0$ pour tout $n \leqq N$". Deux autres principes de permanence sont très utiles: Le *principe de Cauchy*, "aucun ensemble strictement externe n'est interne", n'est rien d'autre que la définition des ensembles strictement externes, et le *principe de Fehrele*, "aucun halo n'est une galaxie", est une généralisation du lemme de Robinson [8], [6].

proche: Deux parties de $\mathbb{R}^n$ sont dites proches si tout point de l'une est équivalent à au moins un point de l'autre. [infinitely close (for sets)]

Robinson (lemme de): voir "principes de permanence".

S-continue: voir "classe S^0"

saut: voir chapitre 1.

standard: C'est un prédicat indéfini (comme $\in$ en théorie des ensembles), dont l'usage (et le sens) est gouverné par trois axiomes: idéalisation, standardisation, et transfert. Tout ensemble est soit standard, soit non standard. Les ensembles $\mathbb{R}$, $\mathbb{Q}$, $\mathbb{C}$, $[0,1]$, ϕ $(=0)$, $\{\phi\}$ $(=1), \cdots, \sin x$, $e^x, \cdots$, sont standard; un infinitésimal ε $(\neq 0)$, la fonction $x \mapsto \varepsilon x$, l'intervalle $[\omega, \omega+1]$ pour ω grand sont non standard. La *partie standard* d'un réel x, notée ${}^0x$, est l'unique standard (s'il existe, voir "limité") qui est équivalent à x. La partie standard d'une fonction f est l'unique fonction standard, notée ${}^0f$, si elle existe, telle que $f(x) \simeq {}^0f(x)$ pour tout x presque standard.

standardisation: C'est l'un des trois principes fondamentaux de l'analyse non standard (voir idéalisation et transfert). Il peut s'énoncer ainsi: "Pour tout ensemble E (interne ou non), il existe un unique ensemble standard, noté ${}^SE$, ayant précisement les mêmes éléments standard que E". Cet ensemble est appelé le *standardisé* de E. L'existence de la partie standard d'un réel limité et de l'ombre d'une partie découle de ce principe.

strictement externe: voir "externe".

transfert: C'est l'un des trois principes fondamentaux de l'analyse non standard. Il s'énonce ainsi: Pour toute formule standard $F(x,t)$, on a, pour t standard:

$$\forall x\ (x \text{ standard} \Rightarrow F(x,t)) \Leftrightarrow \forall x (F(x,t)).$$

REFERENCES

[1] I. VAN DEN BERG, *Permanence principles in non standard analysis*, A paraître.

[2] I. VAN DEN BERG, ET M. DIENER, *Diverses applications du lemme de Robinson*, C. R. Acad Sci. Paris, 293 (1981), pp. 501–504.

[3] F. DIENER, *Méthode du plan d'observabilité*, Thèse, Strasbourg, 1981.

[4] _____, *Les équations* $\varepsilon\ddot{x}+(x^2-1)\dot{x}^{[s]}+x=a$, Collectanea Mathematica, 24, 1978, pp. 217–247.

[5] _____, *Famille d'équations à cycle limite unique*, C. R. Acad. Sci. Paris, 289 (1979), pp. 571–574.

[6] _____, *Cours d'analyse non standard*, Office des Publications Universitaires, Alger, 1983.

[7] _____, *Equations différentielles surquadratiques, et disparition des sauts*, en préparation.

[8] M. DIENER ET I. VAN DEN BERG, *Halos et galaxies. Une extension du lemme de Robinson*, C. R. Acad. Sci. Paris, 293 (1979), pp. 385–388.

[9] W. ECKHAUS, *Asymptotic Analysis of Singular Perturbations*, North-Holland Studies in Mathematics, vol. 9, Amsterdam, 1979.

[10] C. GODBILLON, *Géométrie différentielle et mécanique analytique*, Hermann, Paris, 1969.

[11] F. A. HOWES, *Singularly perturbed non linear boundary-value problems whose reduced equations have singular points*. Studies Appl. Math., 57 (1977), pp. 137–180.

[12] _____, *Boundary-interior layer interactions in non linear singular perturbation theory*, Memoirs Amer. Math. Soc., 203, 1978.

[13] _____, *Singularly perturbed superquadratic boundary value problems*, Nonlinear Anal., 3 (1979), pp. 175–192.

[14] _____, *Some singularly perturbed superquadratic boundary value problems whose solutions exhibit boundary and shock layer behaviour*, Nonlinear Anal., 4 (1980), pp. 683–698.

[15] N. LEVINSON, *An ordinary différential équation with an interval of stability, a separation point and an interval of instability*, J. Math. Phys. 28 (1949), pp. 215–222.

[16] R. LUTZ ET M. GOZE, *Nonstandard Analysis. A Practical Guide With Applications*, Lecture Notes in Mathematics 881, Springer, Berlin, 1981.

[17] R. LUTZ ET T. SARI, *Application of nonstandard analysis to boundary value problems in singular perturbation theory*, Theory and Applications of Singular Perturbations, W. Eckhaus and E. M. de Jager, eds., Lecture Notes in Mathematics, 942, Springer, Berlin, 1982, pp. 113–135.

[18] E. NELSON, *Internal Set Theory*, Bull. Amer. Math. Soc., 83 (1977), pp. 1165–1198.

[19] R. E. O'MALLEY, *Introduction to Singular Perturbations*, Academic Press, New York, 1974.

[20] G. REEB, *Mathématiques non standard (Essai de vulgarisation)*, Bulletin A.P.M.E.P., 328 (1981), pp. 259–273.

[21] A. ROBINSON, *Nonstandard Analysis*, North-Holland, Amsterdam, 1966.

[22] E. URLACHER, *Oscillations de relaxation et analyse non standard*, Thèse, Strasbourg, 1981.

[23] H. I. VISHIK ET L. A. LYUSTERNIK, *Initial jump for nonlinear differential equations containing a small parameter*, Soviet Math. Dokl., 1 (1960), pp. 719–752.

[24] W. R. WASOW, *Asymptotic Expansions for Ordinary Differential Equations*, Interscience, New York, 1965.

SIAM J. MATH. ANAL.
Vol. 17, No. 3, May 1986

004

ASYMPTOTIC ANALYSIS OF SINGULAR SINGULARLY PERTURBED BOUNDARY VALUE PROBLEMS*

CHRISTIAN SCHMEISER† AND RICHARD WEISS†

Abstract. Singularly perturbed systems ordinary differential equations for which the reduced system has a manifold of solutions are called singular singularly perturbed. Boundary value problems for a general class of such systems are examined. Conditions are derived which ensure the existence of a locally unique solution, which can be approximated by an asymptotic expansion. The main tool for our analysis is the theory of boundary value problems on long intervals.

AMS(MOS) subject classifications. Primary 34B15, 34E15

Key words. ordinary differential equations, boundary value problems, singular perturbations, asymptotic expansions

1. Introduction.

We consider boundary value problems of the form

$$\varepsilon y' = f(y,t,\varepsilon) \qquad 0 \leqq t \leqq 1, \tag{1.1}$$

$$b(y(0),y(1)) = 0, \tag{1.2}$$

where y is an n-vector, ε is small and positive, and f and b are nonlinear mappings.

The asymptotic analysis of (1.1), (1.2) starts with an investigation of the reduced system

$$0 = f(y,t,0). \tag{1.3}$$

If there is a locally unique solution $y = Y(t)$ to (1.3) and the matrix $f_y(Y(t),t,0)$ has just strictly stable and strictly unstable eigenvalues, then (1.1), (1.2) is called a regular singular-perturbation problem. Problems of this type are quite well understood. Combining the results of Vasileva and Butuzov (1973) and Esipova (1975) yields a complete asymptotic analysis, even if equations for "slow" components are added to (1.1).

We want to treat the case of the existence of a solution manifold $y = \phi(\alpha,t)$ of (1.3), with α being an n_0-dimensional parameter. We wish to consider problems satisfying an additional assumption:

The matrix $f_y(\phi(\alpha,t),t,0)$ has an n_0-dimensional null space with the real parts of the remaining $n - n_0$ eigenvalues being bounded away from zero.

This assumption rules out the existence of turning points, highly oscillatory solutions and boundary layer variables different from $\tau = t/\varepsilon$ and $\sigma = (1-t)/\varepsilon$. Note that regular singularly perturbed systems of the form

$$\varepsilon y' = g(y,z,t,\varepsilon), \qquad z' = h(y,z,t,\varepsilon)$$

obviously fit into our theory if the equations for z are multiplied by ε.

* Received by the editors February 28, 1984, and in revised form November 18, 1984. This work was supported by "Österreichischer Fonds zur Förderung der wissenschaftlichen Forschung".

† Institut für Angewandte und Numerische Mathematik, Technische Univsersität Wien, A-1040 Wien, Austria.

Singular singularly perturbed problems have received a significant amount of attention recently. General nonlinear initial value problems satisfying the above assumptions have been treated by O'Malley and Flaherty (1980) and Vasileva and Butuzov (1978). In the work by Vasileva and Butuzov also boundary value problems of a special structure are analysed. To the knowledge of the authors no complete classification of problems violating the above assumptions on the eigenstructure of f_y $(\phi(\alpha,t),t,0)$ is available. One of the possible effects in this case is the occurrence of multiple layers. This has been demonstrated on a linear problem by O'Malley (1979).

The asymptotic analysis of (1.1), (1.2) proceeds in two steps: The construction of a formal approximation of the solution and the proof of validity of this approximation.

The first step is contained in §2, where the method of matched asymptotic expansions is used to construct a sequence of formal approximations. The problem defining the first term in the sequence, i.e. the leading term in the asymptotic expansion, is nonlinear in general if (1.1), (1.2) is nonlinear. A key assumption in this paper is the existence of an isolated solution of this problem. It is then shown that this assumption implies the existence and uniqueness of terms of arbitrary order in the asymptotic expansion, since the linear operator in the equations defining these terms is the linearization of the nonlinear problem defining the leading term.

In the proof of the validity of the asymptotic expansion in §4 the contraction principle is used. Thus it is necessary to obtain stability estimates for the linearization of (1.1), (1.2) at the formal approximation, i.e. we have to obtain estimates for the solutions of linear singular singularly perturbed problems, where the coefficients contain boundary layer terms. For the analysis of these problems the theory of boundary value problems on infinite and "long" intervals is used extensively. It is one of the key points of this paper to stress the applicability of this theory to singularly perturbed problems. Some results of this theory are collected in §3. In the proofs given in the Appendix the methods of de Hoog and Weiss (1980) and Markowich (1982, 1983) are used.

The solution of the problem defining the leading term in the expansion is demonstrated on an example from semiconductor theory in §5. The singular perturbation approach to semiconductor problems was originated by Vasileva, Kardosysoev and Stelmakh (1976). Recently the great importance of semiconductor device simulation caused intensive work on the subject. References to some papers using singular perturbation theory are given in §5.

Another nonlinear example can be found in Schmeiser (1985) where a problem modelling large deflections of a thin beam, suggested by Flaherty and O'Malley (1981), is analyzed in the framework of the theory of this paper.

For a class of singular singularly perturbed boundary value problems considered in Vasileva and Butuzov (1978), contraction mapping techniques are employed to establish the validity of the formal asymptotic expansion. The structure of this class is simple enough to a priorily guarantee the existence and uniqueness of an isolated solution to the problem defining the leading term of the expansion and to allow the reduction of the analysis of the linearized problem to that of a scalar second order equation whose coefficients have boundary layers. The stability results employed for this equation have been previously developed in Vasileva (1972).

2. Asymptotic expansion. We consider problems of the form (1.1), (1.2) which satisfy

Hypothesis H1. Denoting the n_0-dimensional solution manifold of the reduced equation by $\phi(\alpha,t)$, the matrix $\phi_\alpha(\alpha,t)$ has constant rank n_0. The Jacobian

$f_y(\phi(\alpha,t),t,0)$ has n_- strictly stable and n_+ strictly unstable eigenvalues with $n_-+n_+ +n_0=n$, for $0\leqq t\leqq 1$.

Differentiation of

$$f(\phi(\alpha,t),t,0)=0$$

with respect to α implies

$$\bar{f}_y\phi_\alpha(\alpha,t)=0, \tag{2.1}$$

where from now on the bar above partial derivatives of f indicates the argument $(\phi(\alpha,t),t,0)$. Thus $\bar{f}_y$ has an n_0-dimensional null space spanned by the columns of $\phi_\alpha(\alpha,t)$. By adding n_++n_- columns which span the stable and unstable subspaces of $\bar{f}_y$ we obtain a transformation matrix $E(\alpha,t)=(E_\pm(\alpha,t),\phi_\alpha(\alpha,t))$ which block-diagonalizes $\bar{f}_y$:

$$E^{-1}\bar{f}_yE=\Lambda=\begin{pmatrix}\Lambda_- & & \\ & \Lambda_+ & \\ & & 0\end{pmatrix}.$$

Denoting the last n_0 rows of $E^{-1}(\alpha,t)$ by $H(\alpha,t)$ it follows that

$$H(\alpha,t)\phi_\alpha(\alpha,t)=I_{n_0} \tag{2.2}$$

and

$$H(\alpha,t)\bar{f}_y=0, \tag{2.3}$$

where I_r is the $r\times r$ identity matrix.

For solutions $y(t,\varepsilon)$ of (1.1),(1.2) we use the ansatz

$$y(t,\varepsilon)\sim\sum_{i=0}^{\infty}(\bar{y}_i(t)+L_iy(\tau)+R_iy(\sigma))\varepsilon^i, \tag{2.4}$$

where $\tau=t/\varepsilon$, $\sigma=(1-t)/\varepsilon$ and

$$\lim_{\tau\to\infty}L_iy(\tau)=0,\qquad \lim_{\sigma\to\infty}R_iy(\sigma)=0,\qquad i=0,1,\cdots. \tag{2.5}$$

2.1. The construction of the leading term. Substituting (2.4) into (1.1) and setting $\varepsilon=0$ yields

$$\bar{y}_0(t)=\phi(\alpha,t). \tag{2.6}$$

Differential equations determing the as yet unknown parameter α as a function of t are obtained from the relations we get by collecting the terms of order ε in (1.1), i.e.

$$\bar{y}_0'=\bar{f}_y\bar{y}_1+\bar{f}_\varepsilon. \tag{2.7}$$

Using (2.6), (2.7) reads

$$\phi_\alpha\alpha'+\phi_t=\bar{f}_y\bar{y}_1+\bar{f}_\varepsilon. \tag{2.8}$$

Multiplying (2.8) by $H(\alpha,t)$ and using (2.2) and (2.3) yields the relations for α,

$$\alpha'=H\big(\bar{f}_\varepsilon-\phi_t\big),\quad 0\leqq t\leqq 1. \tag{2.9}$$

Note that in practice (2.9) is usually obtained from (2.8) without the explicit use of H by eliminating $\bar{y}_1$ from n_0 equations in (2.8).

The equations for the layer corrections are

$$\frac{dL_0y}{d\tau}=f(\phi(\alpha(0),0)+L_0y,0,0),\quad 0\leqq\tau<\infty,\quad L_0y(\infty)=0, \tag{2.10}$$

and

$$\frac{dR_0y}{d\sigma}=-f(\phi(\alpha(1),1)+R_0y,1,0),\quad 0\leqq\sigma<\infty,\quad R_0y(\infty)=0. \tag{2.11}$$

Equating coefficients of order zero in (1.2) yields

$$b(\phi(\alpha(0),0)+L_0y(0),\phi(\alpha(1),1)+R_0y(0))=0. \tag{2.12}$$

We can now state our second fundamental assumption.

Hypothesis H2. The boundary value problem (2.9)–(2.12) has an isolated solution.

Because of the boundary conditions at infinity in (2.10), (2.11) stable manifolds naturally enter our discussion. There is an extensive literature on the subject of invariant manifolds. Some basic results, which can be found in Kelley (1967), imply the existence of an n_--dimensional stable manifold for (2.10) and an n_+-dimensional stable manifold for (2.11). The boundary conditions at infinity require L_0y and R_0y to be trajectories on these manifolds. Then Lemma 3 in Kelley (1967) implies the estimates

$$\begin{aligned}\|L_0y(\tau)\|&\leqq\text{const}\,e^{-\kappa\tau},\\ \|R_0y(\sigma)\|&\leqq\text{const}\,e^{-\kappa\sigma}\end{aligned} \tag{2.13}$$

with a positive constant κ. Subsequently $\|\cdot\|$ will denote a vector norm or the induced matrix norm.

In general, trying to solve problem (2.9)–(2.12) is quite unpleasant. The differential equations on finite and infinite intervals have to be solved simultaneously because there is a coupling by boundary conditions. In applications, however, some knowledge about the structure of the stable manifolds of (2.10) and (2.11) often enables us to obtain from (2.12) n_0 "reduced" boundary conditions for (2.9) alone which do not contain $L_0y(0)$ and $R_0y(0)$. This is the case in all of the above mentioned applications, where the problems on the interval $[0,1]$ and the problems on the infinite interval $[0,\infty)$ can be solved consecutively.

Hypothesis H2 implies that the linearized system

$$w'=\left[H_\alpha\langle\cdot,\tilde f_\varepsilon-\phi_t\rangle+H\left(\bar f_{y\varepsilon}\phi_\alpha-\phi_{\alpha t}\right)\right]w, \tag{2.14a}$$

$$\frac{du}{d\tau}=f_y(\phi(\alpha(0),0)+L_0y,0,0)(u+\phi_\alpha(\alpha(0),0)w(0)), \tag{2.14b}$$

$$\frac{dv}{d\sigma}=-f_y(\phi(\alpha(1),1)+R_0y,1,0)(v+\phi_\alpha(\alpha(1),1)w(1)), \tag{2.14c}$$

$$b_0(\phi_\alpha(\alpha(0),0)w(0)+u(0))+b_1(\phi_\alpha(\alpha(1),1)w(1)+v(0))=0, \tag{2.14d}$$

$$u(\infty)=0,\qquad v(\infty)=0 \tag{2.15}$$

has only the trivial solution. In (2.14d) b_0 and b_1 are the partial derivatives of b with respect to the first and second argument respectively at $(\phi(\alpha(0),0)+L_0y(0),\ \phi(\alpha(1),1)+R_0y(0))$. For bilinear forms B we use the notation $B\langle\cdot,\cdot\rangle$.

2.2. The construction of higher order terms. (2.8) is a linear equation for the smooth part $\bar y_1$ of the first order term. The coefficient matrix $\bar f_y$ is singular. By (2.9) we have chosen the inhomogeneity $\phi_\alpha\alpha'+\phi_t-\bar f_\varepsilon$ in a way that a solution to (2.8) exists. The

general solution of (2.8) can be written in the form

$$\bar{y}_1(t)=\phi_\alpha(\alpha,t)\beta_1(t)+\bar{y}_{1p}(t),$$

where $\bar{y}_{1p}(t)$ is a particular solution and β_1 is an n_0-dimensional parameter. To determine $\beta_1(t)$ we equate coefficients of ε^2 in (1.1) and obtain

$$\phi_\alpha\beta_1'+\phi_\alpha'\beta_1+\bar{y}_{1p}'=\bar{f}_y\bar{y}_2+\frac{1}{2}\bar{f}_{yy}\langle\bar{y}_1,\bar{y}_1\rangle+\bar{f}_{y\varepsilon}\bar{y}_1+\frac{1}{2}\bar{f}_{\varepsilon\varepsilon}. \tag{2.16}$$

Analogously to the determination of α, multiplying (2.16) by H yields

$$\beta_1'=H\left[\frac{1}{2}\bar{f}_{yy}\langle\bar{y}_1,\bar{y}_1\rangle+\bar{f}_{y\varepsilon}\bar{y}_1-\phi_\alpha'\beta_1\right]+H\left[\frac{1}{2}\bar{f}_{\varepsilon\varepsilon}-\bar{y}_{1p}'\right]. \tag{2.17}$$

Vasileva and Butuzov (1978, p. 68) show that (2.17) is a linear equation for β_1. Since we require the coefficient matrix explicitly, we will reproduce their argument. In (2.17) we write

$$\frac{1}{2}H\bar{f}_{yy}\langle\bar{y}_1,\bar{y}_1\rangle=\frac{1}{2}H\bar{f}_{yy}\langle\phi_\alpha\beta_1,\phi_\alpha\beta_1\rangle+H\bar{f}_{yy}\langle\phi_\alpha\beta_1,\bar{y}_{1p}\rangle+\frac{1}{2}\bar{f}_{yy}\langle\bar{y}_{1p},\bar{y}_{1p}\rangle, \tag{2.18}$$

which follows from the fact that $\bar{f}_{yy}$ is a symmetric bilinear form. Differentiating (2.3) with respect to α yields

$$H_\alpha\langle\cdot,\bar{f}_y\cdot\rangle+H\bar{f}_{yy}\langle\phi_\alpha\cdot,\cdot\rangle=0. \tag{2.19}$$

Multiplying (2.19) with ϕ_α gives

$$H_\alpha\langle\cdot,\bar{f}_y\phi_\alpha\cdot\rangle+H\bar{f}_{yy}\langle\phi_\alpha\cdot,\phi_\alpha\cdot\rangle=0.$$

Due to (2.1) ($\bar{f}_y\phi_\alpha=0$) the first term on the right-hand side of (2.18) disappears and (2.17) can be written in the form

$$\beta_1'=H\left[\bar{f}_{yy}\langle\phi_\alpha\cdot,\bar{y}_{1p}\rangle+\bar{f}_{y\varepsilon}\phi_\alpha-\phi_\alpha'\right]\beta_1+\bar{G}_1, \tag{2.20}$$

where $\bar{G}_1$ depends on α and t only. Next we will show that the coefficient matrix S in (2.20) is equal to that in (2.14a). Obviously

$$S=H\left[\bar{f}_{yy}\langle\phi_\alpha\cdot,\bar{y}_{1p}\rangle-\phi_{\alpha\alpha}\langle\alpha',\cdot\rangle\right]+H\left[\bar{f}_{y\varepsilon}\phi_\alpha-\phi_{\alpha t}\right]. \tag{2.21}$$

Differentiation of (2.2) with respect to α gives

$$H_\alpha\langle\cdot,\phi_\alpha\cdot\rangle+H\phi_{\alpha\alpha}\langle\cdot,\cdot\rangle=0. \tag{2.22}$$

It follows from (2.19), (2.21) and (2.22) that

$$S=H_\alpha\langle\cdot,\phi_\alpha\alpha'-\bar{f}_y\bar{y}_{1p}\rangle+H\left[\bar{f}_{y\varepsilon}\phi_\alpha-\phi_{\alpha t}\right].$$

Using the fact that $\bar{y}_{1p}$ is a particular solution of (2.8) it is now obvious that the coefficient matrices in (2.14a) and (2.20) are the same.

Suppose we have constructed $\bar{y}_0,\cdots,\bar{y}_{n-1}$, $n\geqq 2$. Then we determine $\bar{y}_n$ from an equation of the form

$$\bar{f}_y\bar{y}_n=F_n(\bar{y}_0,\cdots,\bar{y}_{n-1}), \tag{2.23}$$

where the inhomogeneity satisfies the solvability condition $HF_n=0$. Thus the general solution of (2.23) can be written as

$$\bar{y}_n=\phi_\alpha\beta_n+\bar{y}_{np}.$$

Equating coefficients of ε^{n+1} in (1.1) gives

$$\phi_\alpha \beta_n' + \phi_\alpha' \beta_n = \bar{f}_y \bar{y}_{n+1} + \bar{f}_{yy}\langle \bar{y}_1, \bar{y}_n \rangle + \bar{f}_{y\varepsilon} \bar{y}_n + \tilde{G}_n,$$

with $\tilde{G}_n$ depending only on $\bar{y}_0, \cdots, \bar{y}_{n-1}$. Multiplying by H and repeating the argument we used for proving the linearity of the equations for β_1 we get

$$\beta_n' = H\left[\bar{f}_{yy}\langle \bar{y}_{1p}, \phi_\alpha \rangle + \bar{f}_{y\varepsilon}\phi_\alpha - \phi_\alpha'\right]\beta_n + \bar{G}_n, \tag{2.24}$$

which has the same coefficient matrix as (2.20). $\bar{G}_n$, like $\tilde{G}_n$, depends only on $\bar{y}_0, \cdots, \bar{y}_{n-1}$. Suppose we have constructed $\bar{y}_0, \cdots, \bar{y}_{n-1}$; $L_0 y, \cdots, L_{n-1} y$; $R_0 y, \cdots, R_{n-1} y$, where the layer corrections $L_i y$, $R_i y$ satisfy exponential estimates of the type (2.13). Then $L_n y$ and $R_n y$ satisfy

$$\frac{dL_n y}{d\tau} = f_y\big(\phi(\alpha(0),0) + L_0 y, 0, 0\big)\big(L_n y + \phi_\alpha(\alpha(0),0)\beta_n(0)\big) + L_n G, \tag{2.25a}$$

$$\frac{dR_n y}{d\sigma} = -f_y\big(\phi(\alpha(1),1) + R_0 y, 1, 0\big)\big(R_n y + \phi_\alpha(\alpha(1),1)\beta_n(1)\big) + R_n G, \tag{2.25b}$$

$$L_n y(\infty) = 0, \qquad R_n y(\infty) = 0, \tag{2.25c}$$

where $L_n G$ and $R_n G$ depend on the terms in the expansion up to order $n-1$. Besides, $L_n G$ and $R_n G$ satisfy exponential estimates of the type (2.13). (2.25a) and (2.25b) differ from (2.14b), (2.14c) only in the terms $L_n G$ and $R_n G$. Equating coefficients of ε^n in (1.2) yields the boundary conditions

$$b_0\big(\phi_\alpha(\alpha(0),0)\beta_n(0) + L_n y(0)\big) + b_1\big(\phi_\alpha(\alpha(1),1)\beta_n(1) + R_n y(0)\big) = c_n, \tag{2.26}$$

where c_n depends on the same terms as $L_n G$ and $R_n G$. These boundary conditions differ from (2.14d) only in the right-hand side c_n. The unique solvability of the problem (2.24) ((2.20) for $n=1$), (2.25) and (2.26) follows from the fact that (2.14), (2.15) has only the trivial solution. $L_n y$ and $R_n y$ are exponentially decaying functions, which is a consequence of Lemma 3.2 in the next section. Thus, the terms in the formal asymptotic expansion (2.4) can be constructed consecutively up to arbitrary order.

3. BVP's on infinite and "long" intervals. In this section we investigate linear problems of the type

$$y' = A(t)y + g(t), \quad t \geqq 0, \quad y \in R^n, \tag{3.1a}$$

$$y(\infty) = 0, \qquad \text{i.e. } \lim_{t\to\infty} y(t) = 0, \tag{3.1b}$$

$$By(0) = \beta. \tag{3.1c}$$

Also we shall derive results on the well-posedness of problems obtained from (3.1) by cutting the infinite interval at a large T and replacing (3.1b) by an appropriate boundary condition at T.

The proofs of the results we state will be given in the Appendix and follow along the lines of de Hoog and Weiss (1980) and Markowich (1982, 1983). In these papers instead of (3.1b) the weaker condition $y \in C[0,\infty]$, i.e. $\lim_{t\to\infty} y(t)$ exists and is finite, is used. However, in the singular perturbation context, (3.1b) is relevant (cf. condition (2.5)). Thus the results are slightly different, but the methods of proof are similar.

3.1. Problems with constant coefficients. We consider the problem

$$y' = My + g(t), \qquad t \geqq 0, \tag{3.2a}$$

$$y(\infty) = 0, \tag{3.2b}$$

$$By(0) = \beta, \tag{3.2c}$$

where the matrix M has a splitting

$$E^{-1}ME = \Lambda = \begin{pmatrix} \Lambda_- & & \\ & \Lambda_+ & \\ & & 0 \end{pmatrix},$$

with the eigenvalues of the n_--dimensional square matrix Λ_- having negative real parts, the eigenvalues of the n_+-dimensional matrix Λ_+ having positive real parts and the zero matrix having dimension $n_0 = n - n_- - n_+$. We denote the column decomposition of E corresponding to the diagonal blocks of Λ by $E = (E_-, E_+, E_0)$ and the row decomposition of E^{-1} by $(E^{-1})^T = ((E_-^{-1})^T, (E_+^{-1})^T, (E_0^{-1})^T)$.

To satisfy (3.2b), we pose the following restrictions on $g(t)$:

$$g \in C[0,\infty], \quad g(\infty) = 0 \quad \text{and} \quad \|E_0^{-1} g(t)\| = O(t^{-1-\nu}), \quad \nu > 0. \tag{3.3}$$

Regarding the unique solvability of (3.2), we have

THEOREM 3.1. *The boundary value problem* (3.2) *has a unique solution for all* $\beta \in R^{n_-}$ *and for all* $g(t)$ *satisfying* (3.3), *iff the matrix* B *has* n_- *rows and* BE_- *is regular.*

LEMMA 3.1. *Let the assumptions of Theorem* 3.1 *and*

$$\|g(t)\| = O(e^{-\kappa t}), \qquad \kappa > 0, \tag{3.4}$$

be valid. Then the solution $y(t)$ *of* (3.2) *satisfies*

$$\|y(t)\| = O(e^{-\kappa t}), \qquad \kappa > 0. \tag{3.5}$$

(*Henceforth in exponential estimates of the type* (3.4), (3.5) κ *will denote a generic constant.*)

Next we consider a problem on the finite interval $[0, T]$:

$$x_T' = Mx_T + g(t), \qquad 0 \leqq t \leqq T, \tag{3.6a}$$

$$\begin{pmatrix} E_+^{-1} \\ E_0^{-1} \end{pmatrix} x_T(T) = \gamma, \tag{3.6b}$$

$$Bx_T(0) = \beta. \tag{3.6c}$$

THEOREM 3.2. *Let the assumptions of Theorem* 3.1 *be valid. Then* (3.6) *has a unique solution* $x_T(t)$ *for all* $T > 0$ *and for all* $g \in C[0,T]$, $\beta \in R^{n_-}$ *and* $\gamma \in R^{n_+ + n_0}$. x_T *satisfies the estimate*

$$\|x_T\|_{[0,T]} \leqq \text{const}\big(T\|g\|_{[0,T]} + \|\beta\| + \|\gamma\|\big), \tag{3.7}$$

where the norm $\|\cdot\|_{[t_1,t_2]}$ *on the space* $C[t_1,t_2]$ *is defined by*

$$\|f\|_{[t_1,t_2]} := \sup_{s \in [t_1,t_2]} \|f(s)\|.$$

3.2. Problems with variable coefficients. We consider problem (3.1) with

$$A(t)=M+F(t), \quad \text{where } \|F(t)\|=O(e^{-\kappa t}). \tag{3.8}$$

In the Appendix we will show that the general solution of the homogeneous problem (3.1a), (3.1b), with $A(t)$ defined by (3.8), is of the form $y(t)=\Phi_-(t)\eta_-$ with $\eta_-\in R^{n_-}$, where the $n\times n_-$-matrix $\Phi_-(t)$ is defined in the Appendix. We then have

THEOREM 3.3. *The boundary value problem* (3.1) *with* $A(t)$ *given by* (3.8) *has a unique solution for all* $\beta\in R^{n_-}$ *and for all* $g(t)$ *satisfying* (3.3), *iff* B *has* n_- *rows and* $B\Phi_-(0)$ *is regular.*

The analogue to Lemma 3.1 is

LEMMA 3.2. *Let the assumptions of Theorem* 3.3 *and* (3.4) *be valid. Then the solution* $y(t)$ *of* (3.1) *satisfies* (3.5).

Again we consider the "finite" problem

$$x_T'=A(t)x_T+g(t), \qquad 0\leqq t\leqq T, \tag{3.9a}$$

$$\begin{pmatrix} E_+^{-1} \\ E_0^{-1} \end{pmatrix} x_T(T)=\gamma, \tag{3.9b}$$

$$Bx_T(0)=\beta. \tag{3.9c}$$

THEOREM 3.4. *Let the assumptions of Theorem* 3.3 *be valid. Then* (3.9) *has a unique solution* $x_T(t)$ *for* T *big enough and for all* $g\in C[0,T]$, $\beta\in R^{n_-}$ *and* $\gamma\in R^{n_++n_0}$. x_T *satisfies the estimate* $\|x_T\|_{[0,T]}\leqq \text{const}(T\|g\|_{[0,T]}+\|\beta\|+\|\gamma\|)$.

THEOREM 3.5. *There exists a fundamental solution* $\psi_T(t)$ *of* (3.9a) *which satisfies*

$$\|\psi_T(t)-X_T(t)\|=O(e^{-\kappa t}), \tag{3.10}$$

where

$$X_T(t)=E\begin{pmatrix} e^{\Lambda_- t} & & \\ & e^{\Lambda_+(t-T)} & \\ & & I \end{pmatrix}$$

is a fundamental solution of (3.6a).

4. Existence and uniqueness result. Let $Y_i(t,\varepsilon)$ denote the ith partial sum in (2.1), i.e.

$$Y_i(t,\varepsilon)=\sum_{j=0}^{i}\left(\bar y_j(t)+L_j y\left(\frac{t}{\varepsilon}\right)+R_j y\left(\frac{1-t}{\varepsilon}\right)\right)\varepsilon^j.$$

Let the space $C^1[0,1]$ be equipped with the norm $\|\cdot\|_*$ defined by

$$\|y\|_*=\|y\|_{[0,1]}+\varepsilon\|y'\|_{[0,1]}.$$

Then $(C^1[0,1], \|\cdot\|_*)$ is a Banach space. Balls in this space will be denoted by

$$B_\delta(y_0)=\{\, y\in C^1[0,1] \mid \|y-y_0\|_*\leqq\delta\,\}.$$

We now prove the main result of this paper.

THEOREM 4.1. *Let* f *and* b *in* (1.1), (1.2) *be infinitely differentiable and let hypotheses* H1, H2 *be valid. Then there are constants* c, $\varepsilon_0>0$ *such that for* $0<\varepsilon\leqq\varepsilon_0$ *a solution* $y(t,\varepsilon)$ *to* (1.1), (1.2) *exists which is unique in the ball* $B_{c\varepsilon}(Y_1)$ *and satisfies*

$$\|y-Y_i\|_*=O(\varepsilon^{i+1}), \qquad i\geqq 0.$$

Proof. Common methods for proving results like Theorem 4.1 are based on the fixed point theorem for contraction mappings in Banach spaces (cf., e.g. Eckhaus (1979) for general results in perturbation theory and Vasileva and Butuzov (1973), (1978) for applications in singular perturbations). We shall apply a theorem due to Van Harten (1978) stated by Eckhaus (1979, p. 237 f). Using his notation, (1.1), (1.2) is written as

$$L_\varepsilon y = 0. \tag{4.1}$$

In our case L_ε is an operator from the Banach space $(C^1[0,1], \|\cdot\|_*)$ to $(C[0,1]\times R^n, \|\cdot\|_{**})$, where we define the norm $\|\cdot\|_{**}$ as the sum of the supremum norm on $C[0,1]$ and the maximum norm on R^n. Our method of constructing the formal approximation Y_i yields

$$L_\varepsilon Y_i = \rho_i, \tag{4.2}$$

where $\|\rho_i\|_{**} = O(\varepsilon^{i+1})$. Subtracting (4.2) from (4.1) yields the following problem for the remainder term $R_i = y - Y_i$:

$$\hat{L}_\varepsilon R_i := L_\varepsilon(R_i + Y_i) - L_\varepsilon Y_i = -\rho_i. \tag{4.3}$$

Denoting the linearization of L_ε at Y_i by A_ε, we obtain that decomposition

$$\hat{L}_\varepsilon = A_\varepsilon + P_\varepsilon.$$

The above mentioned theorem requires us

a) to obtain an estimate

$$\|A_\varepsilon^{-1}\|_* \leqq \lambda(\varepsilon),$$

b) to prove a Lipschitz condition for P_ε of the form

$$\|P_\varepsilon v_1 - P_\varepsilon v_2\|_{**} \leqq \mu(\varepsilon,\delta)\|v_1 - v_2\|_*$$

for v_1, $v_2 \in B_\delta(0)$, where $\lim_{\delta\to 0}\mu(\varepsilon,\delta) = 0$. Then we have to determine $\bar{\delta} = \bar{\delta}(\varepsilon)$, such that

$$\lambda(\varepsilon)\mu(\varepsilon,\delta) \leqq 1-\gamma, \qquad \gamma \in (0,1)$$

for all $\delta \leqq \bar{\delta}$. If

$$\|\rho_i\|_{**} < \frac{\gamma\bar{\delta}}{\lambda(\varepsilon)}$$

then there is a solution R_i of (4.3) which is unique in the ball $B_{\bar{\delta}}(0)$ and satisfies

$$\|R_i\|_* \leqq \frac{1}{\gamma}\lambda(\varepsilon)\|\rho_i\|_{**}.$$

We will show that in our case

$$\mu(\varepsilon,\delta) = c_1\delta, \tag{4.4}$$

$$\lambda(\varepsilon) = \frac{c_2}{\varepsilon}. \tag{4.5}$$

Thus, we obtain $\bar{\delta} = \varepsilon(1-\gamma)/c_1c_2$ and the following condition for ρ_i

$$\|\rho_i\|_{**} < \frac{\varepsilon^2\gamma(1-\gamma)}{c_1c_2^2}.$$

This certainly holds for $i \geqq 2$ and ε small enough. We then have a solution R_i of (4.3), which is unique in $B_{c\varepsilon}(0)$ with $c = (1-\gamma)/c_1 c_2$ and satisfies

$$\|R_i\|_* \leqq \frac{c_2}{\gamma\varepsilon}\|\rho_i\|_{**} = O(\varepsilon^i).$$

By the argument

$$\|R_{i-1}\|_* \leqq \|R_i\|_* + \|Y_i - Y_{i-1}\|_* = O(\varepsilon^i)$$

the results of Theorem 4.1 are established. It remains to show (4.4) and (4.5): Using a formula for $P_\varepsilon v_1 - P_\varepsilon v_2$ (Eckhaus (1979, p. 239)) (4.4) follows immediately.

To establish (4.5), we have to analyze the linearization of (1.1), (1.2), which reads

$$\begin{aligned} &\varepsilon\hat{y}' = f_y(Y_i, t, \varepsilon)\hat{y} + g(t), \\ &\hat{b}_0(Y_i(0), Y_i(1))\hat{y}(0) + \hat{b}_1(Y_i(0), Y_i(1))\hat{y}(1) = \beta \end{aligned} \tag{4.6}$$

or

$$\begin{aligned} &\varepsilon\hat{y}' = f_y(\phi(t) + \varepsilon\bar{y}_1 + L_0 y + R_0 y, t, \varepsilon)\hat{y} + \omega(\varepsilon^2 + \varepsilon e^{-\kappa\tau} + \varepsilon e^{-\kappa\sigma})\hat{y} + g(t), \\ &b_0\hat{y}(0) + b_1\hat{y}(1) = \omega(\varepsilon)(\hat{y}(0), \hat{y}(1)) + \beta, \end{aligned} \tag{4.7}$$

where $\phi(t)$ stands for $\phi(\alpha(t), t)$, b_0 and b_1 are like in §2 and ω is used generically to denote linear operators satisfying $\|\omega(r(\varepsilon, t))\| = O(r(\varepsilon, t))$. As in Vasileva and Butuzov (1973) we introduce a partition of the interval $[0,1]$ into three parts $[0, t_0]$, $[t_0, t_1]$ and $[t_1, 1]$, where

$$t_0 = -\frac{2}{\kappa}\varepsilon\ln\varepsilon, \qquad t_1 = 1 + \frac{2}{\kappa}\varepsilon\ln\varepsilon,$$

and κ is taken from the estimate (2.13). We define

$$\tau_0 = \frac{t_0}{\varepsilon}, \qquad \sigma_1 = \frac{1 - t_1}{\varepsilon}.$$

With (2.13) we obtain

$$\|L_0 y(\tau)\| = O(\varepsilon^2), \quad \tau \geqq \tau_0, \qquad \|R_0 y(\sigma)\| = O(\varepsilon^2), \qquad \sigma \geqq \sigma_1.$$

Introducing the independent variables τ and σ on the short intervals $[0, t_0]$ and $[t_1, 1]$, we obtain a problem equivalent to (4.7),

$$\text{(4.8a)} \quad \frac{d\hat{y}_1}{d\tau} = f_y(\phi(0) + L_0 y, 0, 0)\hat{y}_1 + \omega(\varepsilon\ln\varepsilon)\hat{y}_1 + g(\varepsilon\tau), \qquad 0 \leqq \tau \leqq \tau_0,$$

$$\text{(4.8b)} \quad \frac{d\hat{y}_2}{d\sigma} = -f_y(\phi(1) + R_0 y, 1, 0)\hat{y}_2 + \omega(\varepsilon\ln\varepsilon)\hat{y}_2 + g(1 - \varepsilon\sigma), \qquad 0 \leqq \sigma \leqq \sigma_1,$$

$$\text{(4.8c)} \quad \varepsilon\hat{y}_3' = f_y(\phi(t) + \varepsilon\bar{y}_1, t, \varepsilon)\hat{y}_3 + \omega(\varepsilon^2)\hat{y}_3) + g(t), \qquad t_0 \leqq t \leqq t_1,$$

$$\text{(4.8d)} \quad b_0\hat{y}_1(0) + b_1\hat{y}_2(0) = \omega(\varepsilon)(\hat{y}_1(0), \hat{y}_2(0) + \beta,$$

$$\text{(4.8e)} \quad \hat{y}_1(\tau_0) = \hat{y}_3(t_0), \qquad \hat{y}_2(\sigma_1) = \hat{y}_3(t_1).$$

The transformation

$$\hat{y}_3 = E\xi, \qquad \text{where } \xi = \begin{pmatrix} \xi_\pm \\ \xi_0 \end{pmatrix},$$

changes (4.8c) to

$$\varepsilon\xi' = \left(E^{-1}f_y(\phi(t)+\varepsilon\bar{y}_1,t,\varepsilon)E-\varepsilon E^{-1}E'\right)\xi+\omega(\varepsilon^2)\xi+E^{-1}g(t). \tag{4.9}$$

From

$$f_y(\phi(t)+\varepsilon\bar{y}_1,t,\varepsilon)=\bar{f}_y+\varepsilon\left(\bar{f}_{yy}\langle\cdot,\bar{y}_1\rangle+\bar{f}_{y\varepsilon}\right)+\omega(\varepsilon^2).$$

It follows on using (2.19) and (2.1) that (4.9) can be written as

$$\varepsilon\xi'_{\pm}=\Lambda_{\pm}\xi_{\pm}+\omega(\varepsilon)\xi+E_{\pm}^{-1}g(t), \tag{4.10a}$$

$$\xi'_0=\left(H\bar{f}_{yy}\langle\phi_\alpha\cdot,\bar{y}_{1p}\rangle+H\bar{f}_{y\varepsilon}\phi_\alpha-H\phi'_\alpha\right)\xi_0 \tag{4.10b}$$

$$+H\left[\bar{f}_{yy}\langle E_{\pm}\cdot,\bar{y}_1\rangle+\bar{f}_{y\varepsilon}E_{\pm}-E'_{\pm}\right]\xi_{\pm}+\omega(\varepsilon)\xi+\frac{1}{\varepsilon}Hg(t).$$

With the decoupling transformation

$$\xi_0=\tilde{\xi}_0+\varepsilon H\left[\bar{f}_{yy}\langle E_{\pm}\cdot,\bar{y}_1\rangle+\bar{f}_{y\varepsilon}E_{\pm}-E'_{\pm}\right]\Lambda_{\pm}^{-1}\xi_{\pm},\qquad \tilde{\xi}=\begin{pmatrix}\xi_{\pm}\\ \tilde{\xi}_0\end{pmatrix}$$

(4.10b) becomes

$$\tilde{\xi}'_0=\left(H\bar{f}_{yy}\langle\phi_\alpha\cdot,\bar{y}_{1p}\rangle+H\bar{f}_{y\varepsilon}\phi_\alpha-H\phi'_\alpha\right)\tilde{\xi}_0+\omega(\varepsilon)\tilde{\xi}+\omega\left(\frac{1}{\varepsilon}\right)g(t). \tag{4.11}$$

Introducing the transformation

$$\begin{aligned}\hat{y}_1&=\mu+\phi_\alpha(0)\tilde{\xi}_0(t_0),\\ \hat{y}_2&=\nu+\phi_\alpha(1)\tilde{\xi}_0(t_1),\end{aligned} \tag{4.12}$$

up to $O(\varepsilon\ln\varepsilon)$ terms, (4.8) takes the form

$$\tilde{\xi}'_0=\left[H_\alpha\langle\cdot,\bar{f}_\varepsilon-\phi_t\rangle+H\left(\bar{f}_{y\varepsilon}-\phi_{\alpha t}\right)\right]\tilde{\xi}_0+k_1(t),\qquad t_0\leqq t\leqq t_1, \tag{4.13a}$$

$$\frac{d\mu}{d\tau}=f_y(\phi(0)+L_0y,0,0)(\mu+\phi_\alpha(0)\tilde{\xi}_0(t_0))+k_2(\tau),\qquad 0\leqq\tau\leqq\tau_0, \tag{4.13b}$$

$$\frac{d\nu}{d\sigma}=-f_y(\phi(1)+R_0y,1,0)(\nu+\phi_\alpha(1)\tilde{\xi}_0(t_1))+k_3(\sigma),\qquad 0\leqq\sigma\leqq\sigma_1, \tag{4.13c}$$

$$b_0(\mu(0)+\phi_\alpha(0)\tilde{\xi}_0(t_0))+b_1(\nu(0)+\phi_\alpha(1)\tilde{\xi}_0(t_1))=\beta_1, \tag{4.13d}$$

$$\varepsilon\xi'_{\pm}=\Lambda_{\pm}\xi_{\pm}+k_4(t),\qquad t_0\leqq t\leqq t_1 \tag{4.14a}$$

$$\begin{aligned}&\mu(\tau_0)+\phi_\alpha(0)\tilde{\xi}_0(t_0)=E(t_0)\tilde{\xi}(t_0)+\beta_2,\ \nu(\sigma_1)+\phi_\alpha(1)\tilde{\xi}_0(t_1)\\ &\qquad=E(t_1)\tilde{\xi}(t_1)+\beta_3.\end{aligned} \tag{4.14b}$$

where we now employ arbitrary inhomogeneities k_i and β_j. The equality of the coefficient matrices in (4.11) and (4.13a) has been proven in §2. Thus, the homogeneous equations (4.13) correspond to (2.14), and the solvability of (2.14), (2.15) will now be employed to conclude unique solvability of (4.13), (4.14). The general solutions of the

differential equations in (4.13), (4.14) have the form

$$
\begin{aligned}
&\mu(\tau)=M(\tau)\eta+P(\tau)\tilde{\xi}_0(t_0)+\mu_p(\tau),\\
&\nu(\sigma)=N(\sigma)\delta+Q(\sigma)\tilde{\xi}_0(t_1)+\nu_p(\sigma),\\
&\xi_{\pm}(t)=G_{\pm}(t)\gamma_{\pm}+\xi_{\pm p}(t),\\
&\tilde{\xi}_0(t)=G_0(t)\gamma_0+\tilde{\xi}_{0p}(t).
\end{aligned}
\tag{4.15}
$$

From Theorem 3.5 we obtain for $M(\tau)$ and $N(\sigma)$,

$$
\left\| M(\tau)-E(0)\begin{pmatrix} \exp(\Lambda_-(0)\tau) & \\ & \exp(\Lambda_+(0)(\tau-\tau_0)) \end{pmatrix}_I \right\| = O(e^{-\kappa\tau}),
\tag{4.16a}
$$

$$
\left\| N(\sigma)-E(1)\begin{pmatrix} \exp(-\Lambda_-(1)(\sigma-\sigma_1)) & \\ & \exp(-\Lambda_+(1)\sigma) \end{pmatrix}_I \right\| = O(e^{-\kappa\sigma}),
\tag{4.16b}
$$

and it follows from the proof of Theorem 3.4 that

$$
\begin{aligned}
&\|\mu_p\|_{[0,\tau_0]}\leqq \text{const}\,\tau_0\|k_2\|_{[0,\tau_0]}=\text{const}\ln\frac{1}{\varepsilon}\|k_2\|_{[0,\tau_0]},\\
&\|\nu_p\|_{[0,\sigma_1]}\leqq \text{const}\,\sigma_1\|k_3\|_{[0,\sigma_1]}=\text{const}\ln\frac{1}{\varepsilon}\|k_3\|_{[0,\sigma_i]}.
\end{aligned}
$$

Due to (2.1), $f_y(\phi(0)+L_0y,0,0)\phi_\alpha(0)$ and $f_y(\phi(1)+R_0y,1,0)\phi_\alpha(1)$ decay exponentially. Lemma 3.2 therefore implies

$$
\|P(\tau)\|=O(e^{-\kappa\tau}),\qquad \|Q(\sigma)\|=O(e^{-\kappa\sigma}).
\tag{4.17}
$$

Standard results yield

$$
G_{\pm}(t)=\begin{pmatrix} \exp\left(\Lambda_-(t_0)\dfrac{t-t_0}{\varepsilon}\right) & \\ & \exp\left(\Lambda_+(t_1)\dfrac{t-t_1}{\varepsilon}\right) \end{pmatrix}+O(\varepsilon),
\tag{4.18a}
$$

$$
\|\xi_{\pm p}\|_{[t_0,t_1]}\leqq \text{const}\,\|k_4\|_{[t_0,t_1]}
\tag{4.18b}
$$

and

$$
\|\tilde{\xi}_{0p}\|_{[t_0,t_1]}\leqq \text{const}\|k_1\|_{[t_0,t_1]}.
$$

Substituting the representation (4.15) into the boundary conditions (4.13d) and (4.14b), we obtain

$$
\begin{aligned}
&b_0[M(0)\eta+P(0)G_0(t_0)\gamma_0+\phi_\alpha(0)G_0(t_0)\gamma_0]\\
&\quad+b_1[N(0)\delta+Q(0)G_0(t_1)\gamma_0+\phi_\alpha(1)G_0(t_1)\gamma_0]=\bar{\beta}_1,\\
&M(\tau_0)\eta+P(\tau_0)G_0(t_0)\gamma_0+\phi_\alpha(0)G_0(t_0)\gamma_0-E_{\pm}(t_0)G_{\pm}(t_0)\gamma_{\pm}-\phi_\alpha(t_0)G_0(t_0)\gamma_0=\bar{\beta}_2,\\
&N(\sigma_1)\delta+Q(\sigma_1)G_0(t_1)\gamma_0+\phi_\alpha(1)G_0(t_1)\gamma_0-E_{\pm}(t_1)G_{\pm}(t_1)\gamma_{\pm}-\phi_\alpha(t_1)G_0(t_1)\gamma_0=\bar{\beta}_3,
\end{aligned}
$$

where the $\bar{\beta}_j$ contain the contributions of the β_j and the particular solutions. Denoting the coefficient matrix of this linear system by $C(\varepsilon)$, the estimates (4.16)–(4.18) imply that $C(\varepsilon)=C(0)+\omega(\varepsilon\ln\varepsilon)$, where

$$C(0)=\left[\begin{array}{c|c|cc|c} b_0M(0) & b_1N(0) & 0 & & S \\ 0\ E_+(0)\ E_0(0) & 0 & -E_-(0) & 0 & 0 \\ 0 & E_-(1)\,0\,E_0(1) & 0 & -E_+(1) & 0 \end{array}\right] \tag{4.19}$$

and

$$S=b_0\left[P(0)+\phi_\alpha(0)\right]G_0(0)+b_1\left[Q(0)+\phi_\alpha(1)\right]G_0(1).$$

Here the ordering of the unknowns is η, δ, $\gamma_\pm$, γ_0, with the partitions $\eta^T=(\eta_-^T,\eta_+^T,\eta_0^T)$, $\delta^T=(\delta_-^T,\delta_+^T,\delta_0^T)$, $\gamma_\pm^T=(\gamma_-^T,\gamma_+^T)$. Permutation of the columns in (4.19), corresponding to the ordering η, δ_+, γ_0, $-\gamma_-$, η_+, η_0, δ_-, $-\gamma_+$, δ_0 yields the matrix

$$\begin{pmatrix} \bar{C} & X_1 & X_2 \\ 0 & E(0) & 0 \\ 0 & 0 & E(1) \end{pmatrix} \tag{4.20}$$

where $\bar{C}$, X_1, X_2 are square matrices. It follows on employing Theorem 3.3 that the unique solvability of (2.14), (2.15) is equivalent to C being nonsingular, which implies that the matrix (4.20) is nonsingular. Thus, $C(\varepsilon)$ is nonsingular as well and has a bounded inverse for ε small enough. This and the estimates on the particular solutions imply that (4.13), (4.14) is uniquely solvable for small ε and that

$$\begin{aligned}&\|\tilde{\xi}\|_{[t_0,t_1]},\ \|\mu\|_{[0,\tau_0]},\ \|\nu\|_{[0,\sigma_1]}\\ &\qquad\leqq \text{const}\left(\sum_{i=1}^{3}\|\beta_i\|+\|k_1\|_{[t_0,t_1]}+\|k_4\|_{[t_0,t_1]}+\ln\frac{1}{\varepsilon}\left(\|k_2\|_{[0,\tau_0]}+\|k_3\|_{[0,\sigma_1]}\right)\right).\end{aligned}$$

This allows the application of a contraction mapping argument to obtain unique solvability of (4.8), which is equivalent to (4.6). We thus have unique solvability of (4.6) for ε small enough, with the estimate

$$\|\hat{y}\|_{[0,1]}\leqq\text{const}\left(\|\beta\|+\varepsilon^{-1}\|g\|_{[0,1]}\right).$$

The differential equation in (4.6) can be used to obtain an estimate on $\varepsilon\|\hat{y}'\|_{[0,1]}$ which finally yields

$$\|\hat{y}\|_*\leqq\text{const}\left(\|\beta\|+\varepsilon^{-1}\|g\|_{[0,1]}\right)\leqq\text{const}\,\varepsilon^{-1}\|(\beta,g)\|_{**}.$$

This establishes (4.5) and completes the proof of the theorem.

5. Example. We consider the fundamental semiconductor device equations for the case of a symmetric $p-n$ junction with piecewise constant doping. The singular perturbation approach for this problem was originated by Vasileva, Kardosysoev and Stelmakh (1976). We assume that recombination-generation effects are negligible and that the total current is kept at a prescribed value J. For the scaling which leads to the formulation as a singular perturbation problem, see also Markowich et al. (1982). The

governing equations are

$$\begin{aligned} \varepsilon\psi' &= E, \\ \varepsilon E' &= n-p-1, \\ \varepsilon n' &= nE+\frac{\varepsilon J}{2}, \\ \varepsilon p' &= -pE-\frac{\varepsilon J}{2}. \end{aligned} \tag{5.1}$$

$$\begin{aligned} &\psi(0)=0, \\ &n(0)=p(0), \\ &p(1)=\frac{1}{2}\left(-1+\sqrt{1+4\delta^4}\right)=p_1, \\ &n(1)=p_1+1. \end{aligned} \tag{5.2}$$

The variables ψ, E, n and p are scaled and proportional to the potential, the electric field, the electron density and the hole density in the device. ε and δ result from the scaling. ε is equal to the Debye length, and is small when the doping is large. Thus (5.1) is singularly perturbed in this situation.

In Vasileva and Stelmakh (1977) and Vasileva and Butuzov (1978) (5.1), (5.2) is considered with $\delta=0$. Smith (1980) treats the case of a symmetric $p-n$ junction where the doping and the given electron and hole current densities are not constant. In all these papers results are proven which are more or less equivalent to the one we will obtain below by an application of the general result Theorem 4.1.

The reduced equations

$$0=\bar{E}, \quad 0=\bar{n}-\bar{p}-1, \quad 0=\bar{n}\bar{E}, \quad 0=-\bar{p}\bar{E} \tag{5.3}$$

have the solution

$$\bar{\psi}=\alpha_1, \quad \bar{E}=0, \quad \bar{n}=\alpha_2+1, \quad \bar{p}=\alpha_2. \tag{5.4}$$

The Jacobian of the right-hand side of (5.1) at the solution of (5.3) is

$$\begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 \\ 0 & \alpha_2+1 & 0 & 0 \\ 0 & -\alpha_2 & 0 & 0 \end{pmatrix}$$

with the eigenvalues $\lambda_{1,2}=0$ and $\lambda_{3,4}=\pm\sqrt{2\alpha_2+1}$. Obviously, the matrix has rank 2. Thus hypothesis H1 is fulfilled with the assumption $\alpha_2 \geqq 0$ which is natural because $\bar{p}=\alpha_2$ denotes a density. As in chapter 2 we find differential equations for α_1 and α_2:

$$\begin{aligned} &\text{a)} \quad \alpha_1' = -\frac{J}{2\alpha_2+1}, \\ &\text{b)} \quad \alpha_2' = -\frac{J}{2(2\alpha_2+1)}. \end{aligned} \tag{5.5}$$

The equations for the left layer terms are

$$\begin{aligned}
&\text{a)} && \frac{dL\psi}{d\tau} = LE,\\
&\text{b)} && \frac{dLE}{d\tau} = Ln - Lp,\\
&\text{c)} && \frac{dLn}{d\tau} = (\alpha_2(0)+1+Ln)\,LE,\\
&\text{d)} && \frac{dLp}{d\tau} = -(\alpha_2(0)+Lp)\,LE.
\end{aligned} \tag{5.6}$$

The zeroth order right layer terms disappear. This is due to the fact that the last condition in (5.2) is consistent with (5.3), and will be proved by showing that there is a unique solution of (5.5), (5.6) satisfying the boundary conditions

$$\begin{aligned}
&\text{a)} && \alpha_1(0)+L\psi(0)=0,\\
&\text{b)} && Ln(0)+1=Lp(0),\\
&\text{c)} && \alpha_2(1)=p_1,\\
&\text{d)} && L\psi(\infty)=LE(\infty)=Lp(\infty)=Ln(\infty)=0.
\end{aligned} \tag{5.7}$$

Introducing $L\psi$ as a new independent variable in (5.6c), (5.6d) and solving the resulting linear equations yields with (5.7d)

$$\begin{aligned}
Ln &= (\alpha_2(0)+1)(e^{L\psi}-1),\\
Lp &= \alpha_2(0)(e^{-L\psi}-1).
\end{aligned} \tag{5.8}$$

Now we substitute (5.8) in (5.6b) and solve this equation again with $L\psi$ as independent variable. We obtain

$$LE = -\sqrt{2}\left[(\alpha_2(0)+1)e^{L\psi}+\alpha_2(0)e^{-L\psi}-L\psi-2\alpha_2(0)-1\right]^{1/2}\operatorname{sgn} L\psi, \tag{5.9}$$

where the sign of LE is determined by the condition that the remaining equation for $L\psi$ has to have decaying solutions. (5.8), (5.9) represent the stable manifold of (5.6). Using (5.8) in (5.7b–c) gives

$$\begin{aligned}
&(\alpha_2(0)+1)e^{-\alpha_1(0)} = \alpha_2(0)e^{\alpha_1(0)},\\
&\alpha_2(1)=p_1.
\end{aligned} \tag{5.10}$$

The solution of (5.5), (5.10) is

$$\begin{aligned}
\alpha_1(t) &= \sqrt{1+4\delta^4+2J(1-t)} - \sqrt{1+4\delta^4+2J} + \ln\frac{1+\sqrt{1+4\delta^4+2J}}{\sqrt{4\delta^4+2J}},\\
\alpha_2(t) &= \frac{1}{2}\left(-1+\sqrt{1+4\delta^4+2J(1-t)}\right).
\end{aligned} \tag{5.11}$$

(5.7a) and (5.11) yield an initial value for $L\psi$:

$$L\psi(0) = -\ln\frac{1+\sqrt{1+4\delta^4+2J}}{\sqrt{4\delta^4+2J}}.$$

The proof of the isolatedness of this solution, i.e. the invertibility of the linearization of (5.5)–(5.7), is now outlined. We denote the linearization of the equation (a, b) by $(a, b)_1$ and proceed as follows: The terminal value problem $(5.5b)_1$, $(5.7c)_1$ has a unique solution. The equations $(5.6c, d)_1$ can be integrated similarly to (5.6c, d), and jointly with $(5.7a, b)_1$ lead to an initial condition for $(5.5a)_1$. It remains to determine a decaying solution of $(5.6a, b)_1$ with an initial value defined by $(5.7a)_1$. Writing $(5.6a, b)_1$ as a scalar second order equation, we can use the results of Fife (1974) on such problems to conclude existence and uniqueness of the desired solution.

Thus Hypothesis H2 has been verified and the validity of the formal approximation follows from Theorem 4.1.

Appendix: Proofs of the Results in §3.

Proof of Theorem 3.1. The general solution of (3.2a) is

$$y(t) = Y(t)\eta + y_p(t),$$

where $Y(t) = Ee^{\Lambda t}$ and

$$y_p(t) = H^\delta g(t) = E \begin{pmatrix} \int_\delta^t e^{\Lambda_-(t-s)} E_-^{-1} g(s)\, ds \\ \int_\infty^t e^{\Lambda_+(t-s)} E_+^{-1} g(s)\, ds \\ \int_\infty^t E_0^{-1} g(s)\, ds \end{pmatrix} \quad \text{with } \delta \geqq 0.$$

LEMMA A.1. *$H^\delta g(\infty) = 0$ holds for $g(t)$ satisfying* (3.3).

Proof.

a) $$\left\| \int_\delta^t e^{\Lambda_-(t-s)} E_-^{-1} g(s) \right\| \leqq \text{const} \int_\delta^t e^{-\kappa(t-s)} \| g(s) \|\, ds$$

$$= \text{const} \left(\int_\delta^{(t+\delta)/2} + \int_{(t+\delta)/2}^t \right) e^{-\kappa(t-s)} \| g(s) \|\, ds$$

$$\leqq \text{const} \left((e^{-\kappa(t-\delta)/2} - e^{-\kappa(t-\delta)}) \| g \|_{[\delta, (t+\delta)/2]} + \| g \|_{[(t+\delta)/2, t]} \right)$$

$$\to 0 \text{ for } t \to \infty$$

b) $$\left\| \int_\infty^t e^{\Lambda_+(t-s)} E_+^{-1} g(s)\, ds \right\| \leqq \text{const} \| g \|_{[t, \infty]} \int_t^\infty e^{\kappa(t-s)}\, ds$$

$$\leqq \text{const} \| g \|_{[t, \infty]} \to 0 \text{ for } t \to \infty$$

c) $$\left\| \int_\infty^t E_0^{-1} g(s)\, ds \right\| \leqq \text{const} \int_\infty^t s^{-1-\nu}\, ds = \text{const}\, t^{-\nu} \to 0 \text{ for } t \to \infty.$$

In $Y(t)$ the components corresponding to zero eigenvalues and eigenvalues with positive real parts do not satisfy (3.2b). Therefore, the general solution of (3.2a),(3.2b) is

$$y(t) = E_- e^{\Lambda_- t} \eta_- + H^\delta g(t). \tag{A.1}$$

Substituting (A.1) into (3.2c) shows the validity of Theorem 3.1.

Proof of Lemma 3.1. The proof is similar to that of Lemma A.1 and therefore omitted.

Proof of Theorem 3.2. The general solution of (3.6a) reads

$$x_T(t) = X_T(t)\eta + H_T^\delta g(t), \tag{A.2}$$

where

$$X_T(t) = E \begin{pmatrix} e^{\Lambda_- t} & & \\ & e^{\Lambda_+(t-T)} & \\ & & I \end{pmatrix}$$

and

$$H_T^\delta g(t) = E \begin{pmatrix} \int_\delta^t e^{\Lambda_-(t-s)} E_-^{-1} g(s)\, ds \\ \int_T^t e^{\Lambda_+(t-s)} E_+^{-1} g(s)\, ds \\ \int_T^t E_0^{-1} g(s)\, ds \end{pmatrix}.$$

Similarly to the proof of Lemma A.1 we derive the estimate

$$\| H_T^\delta g \|_{[0,T]} \leqq \text{const} \| g \|_{[0,T]} \text{ for all } g \in C[0,T]. \tag{A.3}$$

Applying the boundary conditions (3.6b), (3.6c) to (A.2) yields

$$BX_T(0)\eta + BH_T^\delta g(0) = \beta,$$
$$\begin{pmatrix} E_+^{-1} \\ E_0^{-1} \end{pmatrix} X_T(T)\eta + \begin{pmatrix} E_+^{-1} \\ E_0^{-1} \end{pmatrix} H_T^\delta g(T) = \gamma.$$

The coefficient matrix in these equations for η is

$$\begin{pmatrix} BE_- & BE_+ e^{-\Lambda_+ T} & BE_0 \\ 0 & I_{n_+} & 0 \\ 0 & 0 & I_{n_0} \end{pmatrix}.$$

Obviously, this matrix is regular since BE_- is regular, and has a bounded inverse for all $T > 0$. This fact together with (A.3) yields unique solvability and the estimate (3.7).

Proof of Theorem 3.3. For solutions of (3.1a), (3.1b)

$$y(t) = E_- e^{\Lambda_- t}\eta_- + H^\delta F y(t) + H^\delta g(t)$$

must hold, which can be written as

$$(I - H^\delta F)\, y(t) = E_- e^{\Lambda_- t}\eta_- + H^\delta g(t), \tag{A.4}$$

where $I - H^\delta F$ is considered as an operator from the space $A_\delta = \{ f \in C[\delta, \infty] \mid f(\infty) = 0 \}$ to itself.

Lemma A.2. $\| H^\delta F f(t) \| \leqq \text{const}\, e^{-\kappa t} \| f \|_{[\delta,\infty]}$, $t \geqq \delta$, *holds for all* $f \in A_\delta$ (*with the constants independent of* δ).

Proof. The proof is similar to that of Lemma A.1 and is therefore omitted.

It follows from Lemma A.2 that

$$\| H^\delta F \|_{[\delta,\infty]} \leqq \text{const}\, e^{-\kappa\delta}.$$

Hence

$$\|H^{\delta_1}F\|_{[\delta_1,\infty]}\leqq\frac{1}{2} \tag{A.5}$$

for some δ_1 big enough. Thus, the operator $I-H^{\delta_1}F$ on A_{δ_1} is invertible and (A.4) yields

$$y(t)=\Phi_-(t)\eta+\tilde{H}g(t),\ t\geqq\delta_1, \tag{A.6}$$

where $\Phi_-(t)=(I-H^{\delta_1}F)^{-1}E_-e^{\Lambda_- t}$ and $\tilde{H}g(t)=(I-H^{\delta_1}F)^{-1}H^{\delta_1}g(t)$. As solutions of (3.1a) Φ_- and $\tilde{H}g$ can be extended to $[0,\infty)$. Applying the boundary conditions (3.1c) to (A.6) confirms the results of Theorem 3.3.

Proof of Lemma 3.2. The result immediately follows from Lemma 3.1, if we can prove

LEMMA A.3. $\|((I-H^{\delta_1}F)^{-1}-I)f(t)\|\leqq\text{const}\,e^{-\kappa t}$ *holds for all* $f\in A_{\delta_1}$.

Proof. We use the identity $(I-H^{\delta_1}F)^{-1}=\sum_{i=0}^{\infty}(H^{\delta_1}F)^i$. It follows from Lemma A.2 and from (A.5) that

$$\|(H^{\delta_1}F)^i f(t)\|\leqq\text{const}\,e^{-\kappa t}2^{1-i}\|f\|_{[\delta_1,\infty]}\text{ for } i\geqq 1.$$

Thus

$$\|((I-H^{\delta_1}F)^{-1}-I)f(t)\|\leqq\sum_{i=1}^{\infty}\text{const}\,e^{-\kappa t}2^{1-i}\|f\|_{[\delta_1,\infty]}=\text{const}\,e^{-\kappa t}.$$

Proof of Theorem 3.4. Solutions of (3.9a) satsify

$$x_T(t)=X_T(t)\eta+H_T^{\delta}Fx_T(t)+H_T^{\delta}g(t),$$

which implies

$$(I-H_T^{\delta}F)x_T(t)=X_T(t)\eta+H_T^{\delta}g(t),$$

where we consider $I-H_T^{\delta}F$ as an operator from $C[\delta,T]$ to itself. An analogous result to Lemma A.2 is

LEMMA A.4. *For* $f\in C[\delta,T]$ *the estimate*

$$\|H_T^{\delta}Ff(t)\|\leqq\text{const}\,e^{-\kappa t}\|f\|_{[\delta,T]},\qquad t\geqq\delta$$

holds (*with the constants independent of* δ *and* T).

Proof. Again the proof is similar to that of Lemma A.1 and is therefore omitted. Lemma A.4 yields

$$\|H_T^{\delta_2}F\|_{[\delta_2,T]}\leqq\frac{1}{2}$$

for δ_2 big enough and all $T\geqq\delta_2$. Thus, $(I-H_T^{\delta_2}F)$ is invertible and we have

$$x_T(t)=\psi_T(t)\eta+\tilde{H}_Tg(t),\qquad \delta_2\leqq t\leqq T, \tag{A.7}$$

where $\psi_T(t)=(I-H_T^{\delta_2}F)^{-1}X_T(t)$ and $\tilde{H}_Tg(t)=(I-H_T^{\delta_2}F)^{-1}H_T^{\delta_2}g(t)$. ψ_T and $\tilde{H}_Tg$ are defined on $[0,T]$ by continuation. An argument as in the proof of Lemma A.3 yields the estimate

$$\|\psi_T(t)-X_T(t)\|\leqq\text{const}\,e^{-\kappa t}. \tag{A.8}$$

Let ψ_T^- be defined by $\psi_T^-(t)=(I-H_T^{\delta_2}F)^{-1}E_-e^{\Lambda_- t}$ for $t\in[\delta_2,T]$, and by continuation for $t\in[0,\delta_2]$. We will now prove that

$$\lim_{T\to\infty}\|\psi_T^- - \Phi_-\|_{[0,T]}=0. \tag{A.9}$$

We define $Z=\psi_T^- - \Phi_-$ and substitute $\bar{\delta}=\max\{\delta_1,\delta_2\}$ for δ in the definitions of H^δ and H_T^δ. The definitions of ψ_T^- and Φ_- imply that Z is defined on $[0,\bar{\delta}]$ by continuation as a solution of $Z'=A(t)Z$. Thus, it suffices to show that

$$\lim_{T\to\infty}\|Z\|_{[\bar{\delta},T]}=0. \tag{A.10}$$

The definitions of ψ_T^- and Φ_- on $[\bar{\delta},T]$ yield

$$Z=H_T^{\bar{\delta}}F\psi_T^- - H^{\bar{\delta}}F\Phi_- = H_T^{\bar{\delta}}FZ + H_T^{\bar{\delta}}F\phi_- - H^{\bar{\delta}}F\phi_-.$$

Thus

$$Z=\left(I-H_T^{\bar{\delta}}F\right)^{-1}\left(H_T^{\bar{\delta}}F\Phi_- - H^{\bar{\delta}}F\Phi_-\right).$$

This implies (A.10), since obviously

$$\lim_{T\to\infty}\left\|H_T^{\bar{\delta}}F\Phi_- - H^{\bar{\delta}}F\Phi_-\right\|_{[\bar{\delta},T]}=0.$$

Now we apply the boundary conditions (3.9b), (3.9c) to (A.7) and obtain equations for η with the coefficient matrix

$$C(T)=\begin{pmatrix} B\psi_T(0) \\ E_+^{-1}\psi_T(T) \\ E_0^{-1}\psi_T(T)\end{pmatrix}.$$

The relations (A.8) and (A.9) imply that

$$\lim_{T\to\infty}C(T)=\begin{pmatrix} B\Phi_-(0) & D \\ 0 & I_{n_+ + n_0}\end{pmatrix}, \tag{A.11}$$

where D is some rectangular matrix. The matrix $B\Phi_-(0)$ is regular which implies regularity of $C(T)$ and boundedness of $\|C^{-1}(T)\|$ for T big enough. An argument similar to the proof of Lemma A.3 and (A.3) yields the estimate

$$\|\tilde{H}_T g\|_{[0,T]}\leqq \text{const}\, T\|g\|_{[0,T]}.$$

This completes the proof of Theorem 3.4.

Proof of Theorem 3.5. It follows from the proof of Theorem 3.4 that $\psi_T(t)$ in (A.7) is a fundamental solution of (3.9a). The inequality (3.10) has been shown during the proof of Theorem 3.4.

REFERENCES

W. ECKHAUS (1979), *Asymptotic Analysis of Singular Perturbations*, North-Holland, Amsterdam.

V. A. ESIPOVA (1975), *The asymptotic behavior of solutions of the general boundary value problem for singularly perturbed systems of ordinary differential equations of conditionally stable type*, Differential Equations, 11, pp. 1457–1465.

P. C. FIFE (1974), *Semilinear elliptic boundary value problems with small parameters*, Arch. Rat. Mech. and Anal., 29, pp. 1–17.

J. E. FLAHERTY AND R. E. O'MALLEY (1981), *Singularly perturbed boundary value problems for nonlinear systems including a challenging problem for a nonlinear beam*, Proc. Oberwolfach 1981, Lecture Notes in Mathematics 942, Springer-Verlag, Berlin.

A. VAN HARTEN (1978), *Non-linear singular perturbation problems: proofs of correctness of a formal approximation based on a contraction principle in a Banach space*, J. Math. Anal. Appl., 65, pp. 126–168.

F. R. DE HOOG AND R. WEISS (1980), *An approximation theory for boundary value problems on infinite intervals*, Computing, 24, pp. 227–239.

A. KELLEY (1967), *The stable, center-stable, center, center-unstable and unstable manifolds*, Appendix in Transversal Mappings and Flows, by R. Abraham and J. Robbin, Benjamin, New York.

P. A. MARKOWICH (1982), *A theory for the approximation of solutions of boundary value problems on infinite intervals*, this Journal, 13, pp. 484–513.

_____ (1983), *Analysis of boundary value problems on infinite intervals*, this Journal, 14, pp. 11–37.

P. A. MARKOWICH, C. RINGHOFER, S. SELBERHERR AND E. LANGER (1982), *An asymptotic analysis of single-junction semiconductor devices*, MRC Technical Summary Report 2527, Mathematics Research Center, Madison, WI.

R. E. O'MALLEY (1979), *A singularly-perturbed linear boundary value problem*, this Journal, pp. 695–708.

R. E. O'MALLEY AND J. E. FLAHERTY (1980), *Analytical and numerical methods for nonlinear singular singularly-perturbed initial value problems*, SIAM J. Appl. Math., 38, pp. 225–248.

C. SCHMEISER (1985), *Finite deformations of thin beams. Asymptotic analysis by singular perturbation methods*, IMA J. Appl. Math., 34, pp. 155–164.

D. R. SMITH (1980), *On a singularly perturbed boundary value problem arising in the physical theory of semiconductors*, Techn. Univ. München, Techn. Rep. TUM-M8021.

A. B. VASILEVA (1972), *The influence of local perturbations on the solution of a boundary value problem*, Differential Equations, 8, pp. 437–443.

A. B. VASILEVA AND V. F. BUTUZOV (1973), *Asymptotic Expansions of Solutions of Singularly Perturbed Differential Equations*, Nauka, Moscow. (In Russian)

_____ (1978), *Singularly perturbed equations in the critical case*, Moscow State University, translated as MRC-Technical Summary Report 2039, Mathematics Research Center, Madison, WI.

A. B. VASILEVA, A. F. KARDOSYSOEV AND V. G. STELMAKH (1976), *Boundary layers in the theory of $p-n$ junctions*, Physics and Technology of Semiconductors, 10, pp. 1321–1329.

A. B. VASILEVA AND V. G. STELMAKH (1977), *Singularly perturbed systems in the theory of transistors*, USSR Comp. Math. and Math. Phys., 17, pp. 48–58.

SIAM J. MATH. ANAL.
Vol. 17, No. 3, May 1986

005

SECOND ORDER NONLINEAR SINGULAR PERTURBATION PROBLEMS WITH BOUNDARY CONDITIONS OF MIXED TYPE*

JENS LORENZ† AND RICHARD SANDERS‡

Abstract. We study nonlinear turning point problems that admit boundary and/or interior layers at positions that are not determined a priori. Our study differs from previous investigations in that for positive "viscosity" first order derivative terms are allowed in the boundary operator. Under certain conditions, shown in a sense to be sharp, we characterize the viscous limit of such problems and prove that they are identical to those limit solutions obtained from the pure Dirichlet problem.

AMS(MOS) subject classifications. Primary 34E20, 34B15, 34A40

Key words. singular perturbations, turning points, bifurcation, viscosity method, mixed boundary conditions

1. Introduction.

In this paper we study boundary value problems of the form:

$$T_\varepsilon u \equiv -\varepsilon u'' + f(u)' + b(x,u) = 0, \qquad 0 \leqq x \leqq 1, \tag{1.1a}$$

$$R_\varepsilon u \equiv \begin{pmatrix} u(0) - \alpha\varepsilon u'(0) \\ u(1) + \beta\varepsilon u'(1) \end{pmatrix} = \begin{pmatrix} \gamma_0 \\ \gamma_1 \end{pmatrix}, \tag{1.1b}$$

where $\varepsilon > 0$ is destined to vanish. For simplicity we assume that $f \in C^2(\mathbb{R})$ and $b \in C^1([0,1]\times\mathbb{R})$ though weaker smoothness assumptions would be sufficient for most of our results. Throughout we define $' = d/dx$ and $f(u)' = a(u)u'$. We also allow $a(u)$ to vanish at arbitrarily many u values, i.e. turning points. If (1.1b) is of Dirichlet type, that is $\alpha = \beta = 0$, and if

$$b_u(x,u) \geqq \mu > 0 \quad \text{for all } (x,u) \in [0,1]\times\mathbb{R} \tag{1.2}$$

then the following basic problem has been solved: For $\varepsilon > 0$, (1.1) has a unique solution u_ε and for ε tending to zero, these functions u_ε tend boundedly a.e. to a limit function u. The limit function u has bounded variation and is characterized by a variational inequality. See [12], [13]; also see [1] for time dependent problems in many space dimensions and see [6], [9] for a treatment of examples.

When α and β are nonzero, the study of the limit behavior of solutions to (1.1) becomes more involved. The difficulty is essentially due to the fact that u_ε' may be of order $1/\varepsilon$ at either boundary; thus the influence of the additional terms in (1.1b) cannot be neglected as ε tends to zero. We shall consider (1.1) under condition (1.2) and also assume that

$$\alpha \geqq 0, \qquad \beta \geqq 0. \tag{1.3a}$$

* Received by the editors October 26, 1984, and in revised form January 23, 1985.

† Department of Applied Mathematics, California Institute of Technology, Pasadena, California, 91125. The research of this author was supported by the National Science Foundation under grant DMS 83-12264.

‡ Department of Mathematics, University of Southern California, Los Angeles, California, 90007. The research of this author was supported by the National Science Foundation under grant MCS 82-00676.

The main results of this paper are now summarized:

(i) Suppose that α and β satisfy (1.3a) as well as

$$-1/\beta \leqq a(u) \leqq 1/\alpha, \tag{1.3b}$$

for all u in an a priori interval $[m,M]$. The interval $[m,M]$ is determined by the maximum principle developed in §2. Then, (1.1) has a unique solution, u_ε, for all $\varepsilon>0$ and u_ε tends boundedly a.e. to a limit solution of bounded variation. Furthermore, this limit u is independent of α and β and is in fact the same function (a.e.) as the one coming from the Dirichlet problem (i.e. the limit of solutions to (1.1) with $\alpha=\beta=0$).

(ii) If (1.3b) is relaxed, that is if either $\max a(u)>1/\alpha$ or $\min a(u)<-1/\beta$, then (1.1) may have multiple solutions even when $\varepsilon>0$. Two solutions of (1.1) are presented in §3 via a bifurcation argument.

The final paragraphs of this section are devoted to a heuristic argument demonstrating the necessity of (1.3b) for uniqueness. First consider the following specific example:

$$-\varepsilon u''+\left(\tfrac{1}{2}u^2\right)'+u=0, \quad u(0)-\alpha\varepsilon u'(0)=2, \quad u(1)=-2. \tag{1.4}$$

Note that the functions

$$u_0(x)=-x+\text{const.}$$

are outer solutions, i.e. they solve the reduced differential equation. (The singular outer solutions $u\equiv 0$ does not play a role in our argument.) There are (at least) two candidates of global outer solutions, namely the discontinuous function

$$\bar{u}_0(x)=\begin{cases} -x+2, & 0\leqq x<\frac{1}{2}, \\ -x-1, & \frac{1}{2}<x\leqq 1 \end{cases} \tag{1.5}$$

and the continuous function $\hat{u}_0(x)=-x-1$, $0<x\leqq 1$. $\bar{u}_0$ satisfies the boundary conditions up to order ε. In the stretched variable $s=(x-x_0)/\varepsilon$ one obtains the inner differential equation

$$-U''+\left(\tfrac{1}{2}U^2\right)'=0, \qquad '=\frac{d}{ds} \tag{1.6}$$

when neglecting terms of order ε. In order to match the inner solution with an outer solution u_0 at an interior discontinuity, say at x_0, we require that

$$\bar{U}(-\infty)=u_0(x_0-0), \qquad \bar{U}(+\infty)=u_0(x_0+0) \tag{1.7}$$

and when u_0 violates the boundary condition at $x=0$ by 0(1), we require that

$$U(0)-\alpha U'(0)=\gamma_0, \qquad U(+\infty)=u_0(0).$$

Integration of (1.6) yields

$$U'(s)=\tfrac{1}{2}\left(U^2(s)-c_0^2\right).$$

Hence with $c_0=\frac{3}{2}$ the matching condition (1.7) is fulfilled for the specific outer solution $\bar{u}_0$ given in (1.5) which satisfies the boundary conditions up to order ε. Therefore $\bar{u}_0$ is a candidate limit solution for all values of α. (It follows from the results proved below that $\bar{u}_0$ is the *only* limit solution for $0\leqq\alpha\leqq\frac{1}{2}$.) When can we fulfill the matching condition at $x=0$ with $\hat{u}_0(x)=-x-1$? The condition $U(+\infty)=\hat{u}_0(0)$ yields

$$U'(s)=\tfrac{1}{2}\left(U^2(s)-1\right).$$

Therefore, if α is a value for which the equation

$$U(0)-\frac{\alpha}{2}\left(U^2(0)-1\right)=2$$

has a solution $U(0)$ in $(-1,1)$, the matching condition can be satisfied. Thus for large values of α other candidates of limit solutions are obtained.

If we generalize the considerations of this specific example slightly, we see that limit solutions different from the Dirichlet limit can possibly occur if α or β are so large that the functions

$$U\to U-\alpha f(u) \quad \text{or } U\to U+\beta f(U)$$

are not monotonically increasing in the range of U-values of interest. But condition (1.3b) exactly requires monotonicity of the above functions.

As we would like to mention, David Brown, California Institute of Technology, found multiple solutions numerically of the type presented here while this paper was written. He used his code described in [10]. We also like to mention that Howes [5] has considered singularly perturbed problems with non-Dirichlet boundary conditions (1.1b) using the method of upper and lower solutions. His assumptions do not allow for turning points, however.

2. Existence and uniqueness for positive ε. In this section we show the problem (1.1) has a solution given that (1.2) and (1.3a) are satisfied. In addition, we show that this solution is unique given (1.3b). We begin with the maximum principle.

LEMMA 2.1. *Assume* (1.2) *and* (1.3a), *and let* u_ε *denote a solution of* (1.1). *Then*

$$m\leqq u_\varepsilon(x)\leqq M, \tag{2.1}$$

where

$$m=\min\left(\gamma_0,\gamma_1,\min\{c(x)\colon 0\leqq x\leqq 1\}\right),$$

$$M=\max\left(\gamma_0,\gamma_1,\max\{c(x)\colon 0\leqq x\leqq 1\}\right).$$

Above, $c(x)$ *is determined by* $b(x,c(x))=0$. *The function* $c(x)$ *is bounded since* $b(x,0)$ *is bounded and* $b_u(x,u)\geqq\mu>0$.

Proof. Let $d=\max\{u_\varepsilon(x)\colon 0\leqq x\leqq 1\}$, and let $y\in[0,1]$ be such that $u_\varepsilon(y)=d$. If $y=0$ we have $u_\varepsilon'(0)\leqq 0$ and (1.1b) implies that $u_\varepsilon(0)\leqq\gamma_0$. If $0<y<1$, the differential equation (1.1a) shows that $b(y,u_\varepsilon(y))\leqq 0=b(y,c(y))$. Thus using (1.2), we have $d=u_\varepsilon(y)\leqq c(y)$. The other estimates follow in a similar manner.

Existence of a solution of (1.1) follows directly from a generalization of the classical Nagumo theorem (see [7], [14]) which has been observed in [4]. Obviously we can apply [4, Thm. 1] using the constant upper and lower functions identically to M and m, respectively, and obtain:

LEMMA 2.2. *Under the conditions* (1.2) *and* (1.3a), *the differential equation* (1.1a) *with boundary conditions* (1.1b) *has a solution* $u_\varepsilon(x)$ *for any* $\varepsilon>0$ *satisfying the estimate* (2.1).

As remarked earlier, conditions (1.2) and (1.3a) are not sufficient to guarantee uniqueness of solutions to (1.1). However, if (1.3b) is also imposed, a solution of (1.1) is unique. We use L^1 techniques to show this. The L^1 techniques used below motivate the characterization of the limit solution presented in §4. Another uniqueness proof is given in an appendix. The second proof, which uses adjoints and an inverse monotonicity

argument, yields additional information; in particular, it shows that solutions u_ε constitute a smooth branch w.r.t. ε. This justifies continuation techniques with ε as a parameter in numerical computations.

THEOREM 2.3. *Suppose conditions* (1.2), (1.3a) *and* (1.3b) *are satisfied. Then the problem* (1.1) *has a unique solution for any* $\varepsilon > 0$. (*When* $\alpha = 0$ *or* $\beta = 0$ *the corresponding estimate in condition* (1.3b) *is considered satisfied.*)

Proof. The method of proof is standard. Define the sign function $\operatorname{sgn}(s)$ by

$$\operatorname{sgn}(s) = \begin{cases} -1 & s < 0, \\ 0, & s = 0, \\ 1, & s > 0, \end{cases}$$

and denote by $\operatorname{sgn}_\delta(s)$, $\delta > 0$, a smooth nondecreasing function with $\operatorname{sgn}_\delta(s) = \operatorname{sgn}(s)$ when $|s| \geqq \delta$ and $s = 0$, and $(d/ds)\operatorname{sgn}_\delta(s) \leqq 2/\delta$ when $|s| \leqq \delta$. For fixed $\varepsilon > 0$ suppose u and v are two solutions of (1.1) and define $w = u - v$. Applying integration by parts to

$$\int_0^1 \operatorname{sgn}_\delta(w)(T_\varepsilon u - T_\varepsilon v)\, dx = 0,$$

one obtains

$$\begin{aligned} &\operatorname{sgn}_\delta(w)\left(-\varepsilon w' + f(u) - f(v)\right)\Big|_0^1 \\ &\quad - \int_0^1 \operatorname{sgn}_\delta(w)'\left(f(u) - f(v)\right) dx \\ &\quad + \int_0^1 \operatorname{sgn}_\delta(w)\left(b(x,u) - b(x,v)\right) dx \leqq 0. \end{aligned} \tag{2.2}$$

Sending δ to zero and using Lebesgue's dominated convergence theorem, we find that the third term above approaches

$$\int_0^1 |b(x,u) - b(x,v)|\, dx.$$

The second term can be shown to approach zero using Lebesgue's theorem again. With the boundary conditions (1.1b) and condition (1.3b) we find that the first term above approaches a nonnegative value. Therefore, we have that any two solutions of (1.1) must satisfy

$$\mu \int_0^1 |u - v|\, dx \leqq \int_0^1 |b(x,u) - b(x,v)|\, dx \leqq 0,$$

and are necessarily the same.

We later need L^1 continuity with respect to the boundary conditions:

LEMMA 2.4. *Suppose* u_1 *and* u_2 *are two solutions of* (1.1a); u_1 (*resp.* u_2) *satisfying the boundary conditions* (1.1b) *where* α, β *are replaced by* α_1, β_1 (*resp.* α_2, β_2). *Further suppose that* $\alpha_1 > 0$, $\beta_1 > 0$ *satisfy* (1.3b). *We then have for* $\alpha_2 > 0$, $\beta_2 > 0$

$$\int_0^1 |u_1 - u_2|\, dx \leqq \text{const.} \left(\left| \frac{1}{\alpha_1} - \frac{1}{\alpha_2} \right| + \left| \frac{1}{\beta_1} - \frac{1}{\beta_2} \right| \right)$$

where the constant above does not depend on $\varepsilon > 0$, α_1, α_2, β_1, β_2.

Proof. Sending δ to zero in (2.2) and using the boundary conditions, one obtains

$$\operatorname{sgn}(w(1))\left\{\frac{u_1(1)-\gamma_1}{\beta_1}-\frac{u_2(1)-\gamma_1}{\beta_2}+f(u_1(1))-f(u_2(1))\right\}$$

$$-\operatorname{sgn}(w(0))\left\{\frac{\gamma_0-u_1(0)}{\alpha_1}-\frac{\gamma_0-u_2(0)}{\alpha_2}+f(u_1(0))-f(u_2(0))\right\}+\mu\int_0^1|u_1-u_2|\,dx\leqq 0$$

with $w=u_1-u_2$. Now use monotonicity of the functions

$$u\to\frac{u}{\beta_1}+f(u),\qquad U\to\frac{u}{\alpha_1}-f(u)$$

in the a priori domain to obtain the result.

3. Rigorous proof of nonuniqueness for (1.3b) violated. If (1.3b) is violated, we show that bifurcation can—and under special assumptions will—occur for problems where (1.2), (1.3a) holds. We actually construct a problem with a nonsingular $u_\varepsilon(x)$ at which *sufficient* conditions for bifurcation are satisfied. First note that a *necessary* condition for bifurcation at $u_\varepsilon(x)$ is the following: The linear problem

$$\begin{aligned}&(T_\varepsilon' u_\varepsilon)v\equiv-\varepsilon v''-\big(a(u_\varepsilon(x))v\big)'+b_u(x,u_\varepsilon(x))v=0,\\&R_\varepsilon v=0\end{aligned}\tag{3.1}$$

has a nontrivial solution v. The following lemma is crucial:

LEMMA 3.1. *Given $\alpha>0$, $\beta\geqq 0$ and $p_0>1/\alpha$. There is $\varepsilon_0>0$ and a smooth function $p(x)$ with $-1/\beta\leqq p(x)\leqq p_0$ such that the eigenvalue problem*

$$L_\varepsilon v\equiv-\varepsilon v''+\big(p(x)v\big)'=\sigma v,\qquad R_\varepsilon v=0\tag{3.2}$$

has a smallest eigenvalue $\sigma_1=\sigma_1(\varepsilon)$ which is less than 0 for $0<\varepsilon\leqq\varepsilon_0$.

Proof. (i) First assume that $p(x)$ in (3.2) is an arbitrary C^1-function. As is well-known, multiplication of the equation $L_\varepsilon v=\sigma v$ by

$$\exp\left(-\frac{1}{\varepsilon}\int_0^x p(s)\,ds\right)$$

transforms the eigenvalue problem (3.2) into a symmetric problem to which the classical Sturm–Liouville theory applies. Thus all eigenvalues of (3.2) are real, and there is a smallest eigenvalue σ_1. To decide upon the sign of σ_1, it suffices to study the homogeneous equation $L_\varepsilon v=0$, which has the general solution

$$v(x)=c_1\phi_1(x)+c_2\phi_2(x)$$

with

$$\phi_1(x)=\exp(r(x)/\varepsilon),\qquad r(x)=\int_0^x p(s)\,ds,$$

$$\phi_2(x)=\phi_1(x)\int_0^x\exp(-r(y)/\varepsilon)\,dy.$$

If we take $c_1=1$, $c_2=(1/\varepsilon)(1/\alpha-p(0))$ we obtain the function

$$\bar v(x)=\phi_1(x)\left\{1+\frac{1}{\varepsilon}\left(\frac{1}{\alpha}-p(0)\right)\int_0^x\exp(-r(y)/\varepsilon)\,dy\right\}\tag{3.3}$$

which satisfies $L_\varepsilon \bar{v}=0$, $R_\varepsilon^{(0)}\bar{v}=0$, $\bar{v}(0)>0$. Here $R_\varepsilon^{(0)}$ denotes the first component of the boundary operator R_ε. With $R_\varepsilon^{(1)}$ we will denote the second component.

(ii) We claim now that the smallest eigenvalue σ_1 of (3.2) is negative if $\bar{v}(1)<0$. To show this, assume that $\bar{v}(1)<0$ and first assume $\sigma_1>0$. If $\sigma_1>0$ then the pair $(L_\varepsilon, R_\varepsilon)$ must be inverse monotone, i.e. for any $w\in C^2[0,1]$ the implication

$$L_\varepsilon w\geqq 0,\ R_\varepsilon^{(0)}w\geqq 0,\ R_\varepsilon^{(1)}w\geqq 0\Rightarrow w\geqq 0 \tag{3.4}$$

holds. (See e.g. [15, Chap. 1, Thm. 17].) But since $L_\varepsilon\bar{v}=0$, $R_\varepsilon^{(0)}\bar{v}=0$ we find that $v\geqq 0$ if $R_\varepsilon^{(1)}v\geqq 0$ and $\bar{v}\leqq 0$ if $R_\varepsilon^{(1)}v\leqq 0$, a contradiction to $\bar{v}(0)\bar{v}(1)<0$. We also arrive at a contradiction if $\bar{v}(1)<0$, $\sigma_1=0$. If $\sigma_1=0$, then $\bar{v}$ is an eigenfunction to σ_1, but the eigenfunction to the smallest eigenvalue is known to be of one sign. (See e.g. [15, Chap. 1, Thm. 16].)

(iii) To conclude the proof of the lemma, note that we allow for $1/\alpha-p(0)<0$ in (3.3). Thus for appropriate $p(x)$ we obtain $\bar{v}(1)<0$ for $0<\varepsilon\leqq\varepsilon_0$ (e.g., take a function $p(x)$ with $p(0)=p_0$, $-1/\beta\leqq p(x)\leqq p_0$, $\int_0^1 \mathrm{p(x)\,dx}<0$).

Assume now $\alpha>0$, $\beta\geqq 0$, $p_0>1/\alpha$ are given and $p(x)$ and $\varepsilon_0>0$ are determined according to Lemma 3.1. We fix $\varepsilon\in(0,\varepsilon_0]$; thus the eigenvalue problem (3.2) has a smallest eigenvalue $\sigma_1<0$. Now take

$$f(u)=\frac{u^2}{2},\qquad a(u)=u,$$

and set $u_\varepsilon(x):=p(x)$,

$$\begin{aligned} B(x)&:=-\varepsilon u_\varepsilon''(x)+f(u_\varepsilon(x))'-\sigma_1 u_\varepsilon(x),\\ b(x,u)&:=-\sigma_1 u-B(x),\\ \gamma&:=R_\varepsilon u_\varepsilon. \end{aligned}$$

By construction the problem

$$\begin{aligned} T_\varepsilon&\equiv -\varepsilon u''+f(u)'+b(x,u)=0,\\ R_\varepsilon u&=\gamma \end{aligned}$$

has $p(x)=u_\varepsilon(x)$ as a solution, and the linearization at $u_\varepsilon(x)$ is

$$\left(T_\varepsilon' u_\varepsilon\right)v=-\varepsilon v''-\left(a(u_\varepsilon(x))v\right)'-\sigma_1 v=-\varepsilon v''+\left(p(x)v\right)'-\sigma_1 v.$$

Thus by construction the eigenvalue problem

$$\left(T_\varepsilon' u_\varepsilon\right)v=\sigma v,\qquad R_\varepsilon v=0 \tag{3.5}$$

has $\sigma=0$ as smallest eigenvalue. Therefore u_ε is a potential bifurcation point if we introduce a parameter λ in the boundary value problem. To make this precise, we use the following theorem from bifurcation theory (see e.g. [16] and see [2], [8] for earlier versions).

THEOREM 3.2. *Let X, Y be real Banach spaces and let*

$$A:\mathbb{R}\times X\to Y$$

denote a twice continuously differentiable operator. Assume $\bar{z}=(\bar{\lambda},\bar{u})\in\mathbb{R}\times X$ *is a point with* $A(\bar{z})=0$ *at which the following conditions hold*:

(i) $\ker(A'(\bar{z}))=\operatorname{span}\{\eta,\xi\}$ *with linear independent* $\eta,\ \xi\in\mathbb{R}\times X$.

(ii) *There is a continuous linear functional* $\psi\colon Y\to\mathbb{R}$ *with* $\operatorname{range}(A'(\bar{z}))=\ker(\psi)$.

(iii) *For the numbers*

$$\bar{a}=\psi\big(A''(\bar{z})\eta\cdot\eta\big),\quad \bar{b}=\psi\big(A''(\bar{z})\eta\cdot\xi\big),\quad \bar{c}=\psi\big(A''(\bar{z})\xi\cdot\xi\big)$$

the hyperbolicity condition $\bar{a}\bar{c}<\bar{b}^2$ *holds.*

Then the solution set of

$$A(z)=0$$

in a neighborhood of $\bar{z}=(\bar{\lambda},\bar{u})$ *consists of exactly two branches which intersect in* $\bar{z}$. *Especially, given any* $\delta>0$ *there is* λ *with*

$$|\lambda-\bar{\lambda}|<\delta$$

for which the equation

$$A(\lambda,\cdot)=0$$

has at least two solutions.

We now claim

LEMMA 3.3. *The abstract Theorem* 3.2 *can be applied if we take* $X=C^2[0,1]$, $Y=C[0,1]\times\mathbb{R}^2$,

$$A(\lambda,u)=\big(T_\varepsilon u+\lambda(u-u_\varepsilon),R_\varepsilon u-\gamma\big)$$

and $(\bar{\lambda},\bar{u})=(0,u_\varepsilon)$, *where* T_ε, R_ε, u_ε, γ *are constructed as above. Especially, for any* $\delta>0$ *there is* λ *with* $|\lambda|<\delta$ *such that the boundary value problem*

$$-\varepsilon u''+\left(\frac{u^2}{2}\right)'+b(x,u)+\lambda(u-u_\varepsilon)=0,\qquad R_\varepsilon u=\gamma$$

has at least two solutions.

Proof. To check conditions (i), (ii), (iii) of Theorem 3.2 first note:

$$\begin{aligned}
&A_\lambda(\lambda,u)\alpha=\big(\alpha(u-u_\varepsilon),0\big),\\
&A_u(\lambda,u)v=\big(\big(T_\varepsilon'u\big)v+\lambda v,R_\varepsilon v\big),\\
&A_{\lambda\lambda}(\lambda,u)\equiv 0,\\
&A_{\lambda u}(\lambda,u)\alpha\cdot v=A_{u\lambda}(\lambda,u)v\cdot\alpha=(\alpha v,0),\\
&A_{uu}(\lambda,u)v\cdot w=\big(\big(T_\varepsilon''u\big)v\cdot w,0\big),
\end{aligned}$$

for $\lambda,\ \alpha\in\mathbb{R}$, $u,v,w\in C^2[0,1]$. By $r(x)$ and $l(x)$ we denote right and left eigenfunctions of (3.5) to $\sigma=0$, i.e.

$$\begin{aligned}
&\big(T_\varepsilon'u_\varepsilon\big)r=0,\qquad R_\varepsilon r=0,\\
&\big(T_\varepsilon'u_\varepsilon\big)^*l=0,\qquad R_\varepsilon^*l=0.
\end{aligned}$$

One can assume

$$r(x)>0,\quad l(x)>0\quad\text{for } 0<x<1$$

since by construction $\sigma=0$ is the smallest eigenvalue of (3.3).

Determination of $\ker A'(0,u_\varepsilon)=:N$: Since $A'(0,u_\varepsilon)(\alpha,v)=((T_\varepsilon'u_\varepsilon)v,R_\varepsilon v)$ it follows that

$$\eta=(0,r), \qquad \xi=(1,0)$$

constitute a basis of N.

Determination of range $(A'(0,u_\varepsilon)):=R$: Let $(c,\beta)\in C\times\mathbb{R}^2$. If $(c,\beta)\in R$ there is $v\in C^2$ with

$$\left(T_\varepsilon'u_\varepsilon\right)v=c, \qquad R_\varepsilon v=\beta.$$

Now choose for $\beta\in\mathbb{R}^2$ a function $w_\beta(x)\in C^2$ with

$$R_\varepsilon w_\beta=\beta$$

such that $\beta\to w_\beta$ is linear. Then

$$R_\varepsilon(v-w_\beta)=0$$

and therefore

$$\begin{aligned}0&=\int_0^1\left\{\left(T_\varepsilon'u_\varepsilon\right)^*l\right\}(v-w_\beta)\,dx\\&=\int_0^1 l\left\{\left(T_\varepsilon'u_\varepsilon\right)(v-w_\beta)\right\}dx=\int_0^1\left\{lc-l\left(T_\varepsilon'u_\varepsilon\right)w_\beta\right\}dx:=\psi(c,\beta).\end{aligned}$$

With the linear functional ψ: $Y\to\mathbb{R}$ defined by the last equation we have shown $R\subset\ker\psi$. But since $\dim N=2$ it follows from Fredholm index theory that $\operatorname{codim}R=1$, and thus $R=\ker\psi$. The hyperbolicity condition: In general holds

$$\begin{aligned}A''(\lambda,u)(\alpha_1,v_1)\cdot(\alpha_2,v_2)=&A_{\lambda\lambda}(\lambda,u)\alpha_1\cdot\alpha_2+A_{\lambda u}(\lambda,u)\alpha_1\cdot v_2\\&+A_{u\lambda}(\lambda,u)v_1\cdot\alpha_2+A_{uu}(\lambda,u)v_1\cdot v_2.\end{aligned}$$

This specializes to

$$\begin{aligned}\bar B&:=A''(0,u_\varepsilon)(0,r)\cdot(1,0)=A_{u\lambda}(0,u_\varepsilon)r\cdot 1=(r,0),\\\bar C&:=A''(0,u_\varepsilon)(1,0)\cdot(1,0)=0.\end{aligned}$$

Therefore

$$\bar b=\psi(\bar B)=\int_0^1 l(x)r(x)\,dx>0$$

and $\bar c=\psi(\bar C)=0$, and thus $\bar a\bar c<\bar b^2$.

Remark. Our construction shows that the condition (1.3b) which guarantees uniqueness in Theorem 2.3, cannot be relaxed, in general. Note that $p_0>1/\alpha$ was arbitrary, and a similar construction can be made if $-1/\beta\leqq a(u)$ is relaxed.

4. The characterization of $\lim_\varepsilon u_\varepsilon$. The question of convergence of u_ε as ε tends to zero is addressed in this section. We show below, given the conditions (1.2), (1.3a) and (1.3b), that as ε tends to zero the solutions to (1.1) converge in L^1 to a limit function. This limit function is characterized by a well-known integral inequality, the so-called entropy inequality as developed by Kruzkov [11], in the study of scalar conservation laws. As will be seen, the Kruzkov theory plays a key role in our analysis. We begin with a simple estimate.

LEMMA 4.1. *Given* (1.2) *and* (1.3a), *we have that any solution of* (1.1) *must satisfy*

$$\int_0^1 \left| \frac{du_\varepsilon}{dx} \right| dx \leqq \text{const.},$$

where the constant above does not depend on $\varepsilon > 0$.

Proof. Differentiating (1.1a) and following the notation of Theorem 2.3, we find that

$$0 = \int_0^1 \operatorname{sgn}_\delta(u_\varepsilon') \cdot \left(-\varepsilon u_\varepsilon''' + f(u_\varepsilon)'' + b(x,u_\varepsilon)' \right) dx.$$

Integrating by parts and using the differential equation (1.1a) in the boundary terms obtained from the integration by parts, we have

$$0 = -\operatorname{sgn}_\delta(u_\varepsilon') b(x,u_\varepsilon)\Big|_0^1 + \int_0^1 \left(\frac{d}{dx} \operatorname{sgn}_\delta(u_\varepsilon') \right) \cdot \left(\varepsilon u_\varepsilon'' - f(u_\varepsilon)' \right) dx$$

$$+ \int_0^1 \operatorname{sgn}_\delta(u_\varepsilon') \cdot \left(b_x(x,u_\varepsilon) + b_u(x,u_\varepsilon) u_\varepsilon' \right) dx.$$

As in Theorem 2.3, sending δ to zero gives us that

$$\int_0^1 b_u(x,u_\varepsilon) |u_\varepsilon'| \, dx \leqq |b(1,u_\varepsilon(1))| + |b(0,u_\varepsilon(0))| + \int_0^1 |b_x(x,u_\varepsilon)| \, dx,$$

from which hypothesis (1.2) and the maximum principle, Lemma 2.1, make the final result obvious.

The importance of the estimate obtained above is the following. As is well-known, any sequence of functions having uniformly bounded variation is sequentially compact in the L^1 topology; see e.g. [3, Chap. IV, 8.20]. Lemma 4.1 tells us that $\{u_\varepsilon\}_{\varepsilon > 0}$ has variation which is bounded uniformly for $\varepsilon > 0$. We therefore have a function, say u, along with a sequence ε tending to zero, such that $u_\varepsilon \to u$ in L^1. What remains to be shown is that u is the unique limit point of u_ε with ε tending to zero.

For the moment, we shall assume a strict version of condition (1.3b). We assume that α and β in the boundary conditions (1.1b) satisfy

$$-1/\beta < a(u) < 1/\alpha, \tag{4.1}$$

for all u in the a priori domain $[m,M]$. It will be seen by virtue of the continuity estimate, Lemma 2.4, that (4.1) is an adequate assumption to prove the main result stated in the introduction. This point is discussed further in the proof of Theorem 4.6.

With condition (4.1) in place over (1.3b), we now have a lemma concerning the behavior of u_ε at $x = 0$ and at $x = 1$.

LEMMA 4.2. *Given* (1.2), (1.3a) *and* (4.1), *and assume that* u_ε *tends in* L^1 *to a function* u *of bounded variation for some sequence* ε_n *tending to zero. We then have, for the same sequence* ε_n, *that*

(i) $\lim u_\varepsilon(0)$ *and* $\lim u_\varepsilon(1)$ *exist, and*

(ii)

$$\lim \varepsilon u_\varepsilon'(0) = \lim f(u_\varepsilon(0)) - f(\gamma u(0)),$$
$$\lim \varepsilon u_\varepsilon'(1) = \lim f(u_\varepsilon(1)) - f(\gamma u(1)).$$

$\gamma u(0)$ *and* $\gamma u(1)$ *denote the boundary traces of the BV function* u.

Proof. To prove (i) at $x=1$ we may assume that $\beta>0$ in (1.1b); otherwise there is nothing to prove. Define ρ to be a smooth nondecreasing function with $\rho(0)=0$ and $\rho(1)=1$. Multiplying (1.1a) with ρ and integrating by parts, we find that

$$\varepsilon u_\varepsilon'(1)-f(u_\varepsilon(1))=\int_0^1 \left(\rho'\left(\varepsilon u_\varepsilon'-f(u_\varepsilon)\right)+\rho b(x,u_\varepsilon)\right)dx. \tag{4.2}$$

(1.1b) gives us that $\varepsilon u_\varepsilon'(1)=(\gamma_1-u_\varepsilon(1))/\beta$. Therefore, (4.2) may be written as

$$u_\varepsilon(1)+\beta f(u_\varepsilon(1))=\gamma_1-\beta\int_0^1 \left(\rho'\left(\varepsilon u_\varepsilon'-f(u_\varepsilon)\right)+\rho b(x,u_\varepsilon)\right)dx.$$

Condition (4.1) implies that the left-hand side of this identity is a strictly increasing function. The right-hand side has a limit by virtue of the hypotheses of the lemma. So we have $\lim u_\varepsilon(1)$ exists and similarly $\lim u_\varepsilon(0)$ exists.

To prove (ii), we return to (4.2). Sending ε to zero, we find that

$$\lim \varepsilon u_\varepsilon'(1)=\lim f(u_\varepsilon(1))+\int_0^1 \left(-\rho' f(u)+\rho b(x,u)\right)dx,$$

where $u=\lim u_\varepsilon$ in L^1. Since u has bounded variation, a simple exercise would show that

$$\lim_{\delta\to 0} 1/\delta\int_{1-\delta}^1 f(u)\,dx=f\Big(\lim_{x\uparrow 1} u(x)\Big)\equiv f(\gamma u(1)).$$

Setting

$$\rho(x)=\begin{cases}1+(x-1)/\delta, & 1-\delta\leqq x\leqq 1,\\ 0, & x\leqq 1-\delta,\end{cases}$$

we have

$$\begin{aligned}\lim \varepsilon u_\varepsilon'(1)&=\lim f(u_\varepsilon(1))+\lim_{\delta\to 0}\left[-1/\delta\int_{1-\delta}^1 f(u)\,dx+O(\delta)\right]\\ &=\lim f(u_\varepsilon(1))-f(\gamma u(1)),\end{aligned}$$

which completes the proof of the lemma.

Next we obtain an integral inequality that every limit function of u_ε, with ε tending to zero, satisfies. This inequality is frequently referred to as an entropy inequality in the literature. A significant implication of the entropy inequality derived below is that only one BV function can satisfy it; see Proposition 4.5. This fact combined with the result of Lemma 4.1 implies that u_ε converges to a limit u independent of any particular sequence of ε-values tending to zero. The entropy inequality therefore characterizes the limit function of u_ε.

LEMMA 4.3. *Given* (1.2), (1.3a) *and* (4.1), *we have that any* L^1 *limit of* u_ε, *say* u, *must satisfy*

$$\begin{aligned}-\int_0^1 \operatorname{sgn}(u-k)\big((f(u)-f(k))\phi_x-b(x,u)\phi\big)\,dx&\\ \leqq \operatorname{sgn}(u(1)-k)\big(f(k)-f(\gamma u(1))\big)\phi(1)&\\ -\operatorname{sgn}(u(0)-k)\big(f(k)-f(\gamma u(0))\big)\phi(0),&\end{aligned}$$

where k *is any real number and* ϕ *is any smooth nonnegative function. Above,* $u(0)=\lim u_\varepsilon(0)$, $u(1)=\lim u_\varepsilon(1)$ *and they are in general not equal to* $\gamma u(0)$, $\gamma u(1)$.

Proof. Multiply the identity

$$-\varepsilon u_\varepsilon'' + \{f(u_\varepsilon) - f(k)\}' + b(x, u_\varepsilon) = 0$$

by $\operatorname{sgn}_\delta(u_\varepsilon - k)\phi$, integrate from zero to one using integration by parts, and send δ to zero. Then obtain

$$\begin{aligned} &-\int_0^1 \operatorname{sgn}(u_\varepsilon - k)\big((f(u_\varepsilon) - f(k))\phi_x - b(x, u_\varepsilon)\phi\big)\,dx \\ &\quad \leqq -\operatorname{sgn}(u_\varepsilon - k)(f(u_\varepsilon) - f(k))\phi\big|_0^1 + \operatorname{sgn}(u_\varepsilon - k)\varepsilon u_\varepsilon'\phi\big|_0^1 \\ &\qquad -\varepsilon\int_0^1 \operatorname{sgn}(u_\varepsilon - k)u_\varepsilon'\phi_x\,dx. \end{aligned}$$

The last term on the right-hand side above goes to zero with ε. For k different from $u(0)$ and $u(1)$, the second result of Lemma 4.2 inserted into the $\varepsilon u_\varepsilon'$-term above would complete the proof. Taking limits with k approaching $u(0)$ or $u(1)$ from above and below will establish the entropy inequality also in the exceptional cases; that is, those cases in which the discontinuity of the sign function must be taken into account.

The limit function u coming from (1.1) with mixed boundary conditions also satisfies the entropy inequality satisfied by limit solutions of the pure Dirichlet problem (i.e. $u_\varepsilon(0) = \gamma_0$ and $u_\varepsilon(1) = \gamma_1$). This fact is fundamental in proving the main result of this section. With this in mind, we state:

LEMMA 4.4. *With the same assumptions of Lemma* 4.3, *we have that u satisfies*

(4.3)

$$\begin{aligned} &-\int_0^1 \operatorname{sgn}(u - k)\big((f(u) - f(k))\phi_x - b(x, u)\phi\big)\,dx \\ &\quad \leqq \operatorname{sgn}(\gamma_1 - k)(f(k) - f(\gamma u(1)))\phi(1) - \operatorname{sgn}(\gamma_0 - k)(f(k) - f(\gamma u(0)))\phi(0), \end{aligned}$$

for all real numbers k and any smooth nonnegative ϕ.

Proof. The proof of the lemma is immediate once that

$$\operatorname{sgn}(u(1) - k)(f(k) - f(\gamma u(1))) \leqq \operatorname{sgn}(\gamma_1 - k)(f(k) - f(\gamma u(1))) \tag{4.4}$$

and a similar inequality at 0 is shown. To prove (4.4), first note that (4.4) is nontrivial only if $\beta > 0$ and $\gamma_1 \leqq k \leqq u(1)$ or $u(1) \leqq k \leqq \gamma_1$. For definiteness let

$$\gamma_1 \leqq k \leqq u(1). \tag{4.5}$$

Using (ii) from Lemma 4.2 and the boundary condition (1.1b), we have that

$$u(1) + \beta f(u(1)) = \gamma_1 + \beta f(\gamma u(1)).$$

Since the function $v \to v + \beta f(v)$ increases, we find with (4.5) that

$$\begin{aligned} k + \beta f(k) &\leqq u(1) + \beta f(u(1)) \\ &= \gamma_1 + \beta f(\gamma u(1)) \leqq k + \beta f(\gamma u(1)); \end{aligned}$$

thus $f(k) \leqq f(\gamma u(1))$. This shows (4.4).

We next prove that only one L^1 limit function of u_ε can satisfy the entropy inequality of Lemma 4.4. This result is not new; see [1], [12] for example; however, for completeness we shall sketch the proof.

PROPOSITION 4.5. *Suppose u_1 and u_2 are two BV functions that satisfy the entropy inequality of Lemma* 4.4. *Further let $b_u \geqq 0$. Then*

$$\int_0^1 |b(x,u_1)-b(x,u_2)|\,dx=0.$$

Proof. Take a nonnegative test function ϕ with $\phi(y)=\phi(-y)$ for all y. Let $u=u_1(x)$ and set $k=u_2(x')$ and $\phi=\phi(x-x')$ in (4.3). Integrate the result with respect to the variable x'. Reversing the roles of u_1 and u_2, x and x', and adding the two inequalities together yields

(4.6)

$$\int_0^1\int_0^1 \operatorname{sgn}(u_1(x)-u_2(x'))\{b(x,u_1(x))-b(x',u_2(x'))\}\phi(x-x')\,dx'\,dx$$

$$\leqq \sum_{i=0,1}\eta_i\int_0^1 (f(u_2)-f(\gamma u_1(i)))\operatorname{sgn}(\gamma_i-u_2)\phi(i-x')\,dx'$$

$$+\sum_{i=0,1}\eta_i\int_0^1 (f(u_1)-f(\gamma u_2(i)))\operatorname{sgn}(\gamma_i-u_1)\phi(x-i)\,dx.$$

Here we used the notation $\eta_i=(-1)^{i+1}$. Now, letting $\phi(x)$ approach the delta function, we have that the above double integral tends to

$$\int_0^1 |b(x,u_1)-b(x,u_2)|\,dx.$$

The right-hand side of (4.6) can be treated as follows. Consider, e.g., the first term for $i=1$ and find

$$\int_0^1 [f(u_2)-f(\gamma u_1(1))]\operatorname{sgn}(\gamma_1-u_2)\phi(1-x)\,dx$$

$$\leqq\int_0^1 [f(u_2)-f(\gamma_1)]\operatorname{sgn}(\gamma_1-u_2)\phi(1-x)\,dx$$

$$+\int_0^1 |f(\gamma_1)-f(\gamma u_1(1))|\phi(1-x)\,dx.$$

For $\phi(x)\to\delta(x)$ the right-hand side of this inequality approaches

$$\frac{1}{2}[f(\gamma u_2(1))-f(\gamma_1)]\operatorname{sgn}(\gamma_1-\gamma u_2(1))+\frac{1}{2}|f(\gamma_1)-f(\gamma u_1(1))|.$$

The other terms on the right-hand side of (4.6) can be treated similarly and one finds the estimate

$$\int_0^1 |b(x,u_1)-b(x,u_2)|\,dx$$

$$\leqq\frac{1}{2}\sum_{i=0,1}\eta_i\{[f(\gamma u_2(i))-f(\gamma_i)]\operatorname{sgn}(\gamma_i-\gamma u_2(i))$$

$$+[f(\gamma u_1(i))-f(\gamma_i)]\operatorname{sgn}(\gamma_i-\gamma u_1(i))$$

$$+\eta_i|f(\gamma_i)-f(\gamma u_2(i))|+\eta_i|f(\gamma_i)-f(\gamma u_1(i))|\}.$$

It thus remains to be shown for $u=u_1$ and $u=u_2$:

$$\eta_i \operatorname{sgn}(\gamma_i-\gamma u(i))[f(\gamma u(i))-f(\gamma_i)] \leqq 0.$$

But this inequality follows from (4.3) if one takes there $k=\gamma_i$ and chooses ϕ as in the proof of Lemma 4.4.

We finally are ready to state and prove the main result of this section.

THEOREM 4.6. *Suppose u_ε is a sequence of solutions to* (1.1) *with ε tending to zero. Further suppose that the differential equation satisfies conditions* (1.2), (1.3a) *and* (1.3b). *Then, u_ε tends boundedly a.e. to a limit function and this limit function is characterized by the entropy inequality* (4.3). *Furthermore, the limit function is independent of α and β and is in fact the same a.e. as the limit function coming from the Dirichlet problem.*

Proof. First, the combined results of this section prove the theorem in the case when the strict condition (4.1) is in place over (1.3b). That is, we have shown that every sequence of u_ε, ε tending to zero, has a convergent subsequence to a unique limit function which is independent of α or β. What remains to be shown is that the theorem remains valid when (4.1) is weakened to (1.3b). To show this, we consider a second family of solutions, $u_{\delta,\varepsilon}$, to a modified problem:

$$\begin{aligned} &T_\varepsilon u_{\delta,\varepsilon}=0,\\ &u_{\delta,\varepsilon}(0)-\alpha(1-\delta)\varepsilon u'_{\delta,\varepsilon}(0)=\gamma_0,\\ &u_{\delta,\varepsilon}(1)+\beta(1-\delta)\varepsilon u'_{\delta,\varepsilon}(1)=\gamma_1, \end{aligned}$$

where we may assume that $\alpha>0$ and $\beta>0$. If the boundary conditions above satisfy (1.3b) with $\delta=0$, then they satisfy (4.1) with $0<\delta<1$. So we have

$$\|u_\varepsilon-u\|_1 \leqq \|u_\varepsilon-u_{\delta,\varepsilon}\|_1+\|u_{\delta,\varepsilon}-u\|_1$$

where u is the limit of solutions to the Dirichlet problem.

Now, let η be an arbitrary positive number. The continuity estimate of Lemma 2.4 allows us to fix $\delta>0$ sufficiently small such that

$$\|u_\varepsilon-u_{\delta,\varepsilon}\|_1<\eta/2$$

for all $\varepsilon>0$. Since $u_{\delta,\varepsilon}$ satisfies (4.1), we may choose ε sufficiently small such that

$$\|u_{\delta,\varepsilon}-u\|_1<\eta/2,$$

thus completing the proof of the theorem.

5. Appendix: Alternate proof of uniqueness using inverse-monotonicity. Given the problem (1.1) with (1.2), (1.3a, b) is satisfied. The linearization of T_ε at any function $u(x)\in C^2$ reads

$$(T'_\varepsilon u)v=-\varepsilon v''+(a(u(x))v)'+b_u(x,u(x))v.$$

Thus we study linear operators

$$L_\varepsilon v=-\varepsilon v''+(p(x)v)'+q(x)v. \tag{5.1}$$

The following lemma is the key of the uniqueness argument:

LEMMA 5.1. *Let L_ε be given by* (5.1) *and assume $q(x)>0$ in* $[0,1]$ *as well as*

$$-\frac{1}{\beta}\leqq p(x)\leqq\frac{1}{\alpha} \quad \text{in } [0,1]$$

where $\alpha \geqq 0$, $\beta \geqq 0$ *determine* R_ε *as in* (1.1b). *Then* $(L_\varepsilon, R_\varepsilon)$ *is inverse-monotone, i.e.* (3.4) *holds. Especially, the homogeneous equation* $L_\varepsilon v = 0$, $R_\varepsilon v = 0$ *has only the trivial solution.*

Proof. Let $L_\varepsilon^* u = -\varepsilon u'' - p(x)u' + q(x)u$. Using the notation

$$(c_1, c_2)_2 = \int_0^1 c_1(x) c_2(x)\, dx$$

we have for all $u, v \in C^2$:

$$(L_\varepsilon v, u)_2 = (v, L_\varepsilon^* u)_2 + B(u, v)$$

with $B(u,v) = \{-\varepsilon(v'u - vu') + pvu\}\,|_0^1$. If $R_\varepsilon v = 0$ one has for all $u \in C^2$:

$$B(u,v) = \varepsilon v'(0)\{(1 - \alpha p(0))u(0) - \alpha\varepsilon u'(0)\}$$
$$-\varepsilon v'(1)\{(1 + \beta p(1))u(1) + \beta\varepsilon u'(1)\}.$$

Thus we define the adjoint boundary operator

$$R_\varepsilon^* u = \begin{pmatrix} (1 - \alpha p(0))u(0) - \alpha\varepsilon u'(0) \\ (1 + \beta p(1))u(1) + \beta\varepsilon u'(1) \end{pmatrix}.$$

Thus $R_\varepsilon v = R_\varepsilon^* u = 0$ implies

$$(L_\varepsilon v, u)_2 = (v, L_\varepsilon^* u)_2.$$

Since $q(x) > 0$, $1 - \alpha p(0) \geqq 0$, $1 + \beta p(1) \geqq 0$ we have

$$L_\varepsilon^* \bar{e} > 0, \qquad R_\varepsilon^* \bar{e} \geqq 0$$

with $\bar{e}(x) \equiv 1$; thus $\bar{e}(x)$ is a majorizing function for $(L_\varepsilon^*, R_\varepsilon^*)$. Then it is known that $(L_\varepsilon^*, R_\varepsilon^*)$ has a nonnegative Green's function (e.g. [15]) and the assertion follows. The following theorem is an immediate consequence of Lemma 5.1.

THEOREM 5.2. *For the problem* (1.1) *assume*

$$b_u(x,u) \geqq \mu > 0, \qquad \alpha \geqq 0, \quad \beta \geqq 0$$

and let

$$-\frac{1}{\beta} \leqq a(u) \leqq \frac{1}{\alpha} \quad \text{for } m \leqq u \leqq M$$

where $[m, M]$ *denotes the a priori domain.*

(i) *If* $u(x)$, $v(x) \in C^2$ *satisfy*

$$m \leqq u(x), \qquad v(x) \leqq M,$$

then

$$T_\varepsilon u \leqq T_\varepsilon v, \qquad R_\varepsilon u \leqq R_\varepsilon v$$

implies $u \leqq v$. *Especially,* (1.1) *is uniquely solvable.*

(ii) *At any* $u(x) \in C^2$ *with* $m \leqq u(x) \leqq M$, *the linear problem*

$$(T_\varepsilon' u)v = 0, \qquad R_\varepsilon v = 0$$

only has the trivial solution. As a consequence, if $u_\varepsilon(x)$ *denotes the solution of* (1.1), *the mapping* $\varepsilon \to u_\varepsilon$ *is continuous from* $(0, \infty)$ *into* C^2.

Proof. We only have to show (i). For $0 \leqq s \leqq 1$ let

$$z_s(x) = v(x) + s[u(x) - v(x)].$$

If we set

$$p(x) = \int_0^1 a(z_s(x))\, ds,$$
$$q(x) = \int_0^1 b_u(x, z_s(x))\, ds$$

and define L_ε by (5.1) then

$$T_\varepsilon u - T_\varepsilon v = L_\varepsilon(u - v).$$

The assertion now follows directly from Lemma 5.1.

REFERENCES

[1] C. BARDOS, A.Y. LE ROUX, AND J. C. NEDELEC, *First order quasilinear equations with boundary conditions*, Comm. PDE, 4 (1979), pp. 1017–1034.

[2] M. G. CRANDALL AND P. H. RABINOWITZ, *Bifurcation from simple eigenvalues*, J. Funct. Anal., 8 (1971), pp. 321–340.

[3] N. DUNFORD AND J. SCHWARTZ, *Linear Operators, Part* 1: *General Theory*, Pure and Applied Mathematics, Vol. VII, Interscience, New York, 1958.

[4] J. W. HEIDEL, *A second order nonlinear boundary value problem*, J. Math. Anal. Appl., 48 (1974), pp. 493–503.

[5] F. A. HOWES, *Singular perturbations and differential inequalities*, Mem. Amer. Math. Soc., 168 (1976).

[6] _____, *Boundary-interior layer interactions in nonlinear singular perturbation theory*, Mem. Amer. Math. Soc., 203 (1978).

[7] L. K. JACKSON, *Subfunctions and second order ordinary differential inequalities*, Adv. Math., 2 (1968), pp. 307–363.

[8] H. B. KELLER, *Numerical solution of bifurcation and nonlinear eigenvalue problems*, in Applications of Bifurcation Theory, P. H. Rabinowitz, ed., Academic Press, New York, 1977.

[9] J. KEVORKIAN AND J. D. COLE, *Perturbation Methods in Applied Mathematics*, Springer, New York, 1981.

[10] H. O. KREISS, N. K. NICHOLS, AND D. L. BROWN, *Numerical methods for stiff two-point boundary value problems*, MRC TS Rept. 2599, 1983.

[11] S. N. KRUZKOV, *First order quasi-linear equations in several independent variables*, Math. USSR Sb., 10 (1970), pp. 217–243.

[12] J. LORENZ, *Nonlinear boundary value problems with turning points and properties of different schemes*, in Lecture Notes in Mathematics 942, W. Eckhaus and E. M. de Jager, eds., Springer, Berlin, 1982.

[13] _____, *Analysis of difference schemes for a stationary shock problem*, SIAM J. Numer. Anal., 21 (1984), pp. 1038–1053.

[14] M. NAGUMO, *Über die Differentialgleichung* $y'' = f(x, y, y')$, Proc. Phys. Math. Soc. Japan, 19 (1937), pp. 861–866.

[15] M. H. PROTTER AND H. I. WEINBERGER, *Maximum principles in differential equations*, Prentice-Hall, Englewood Cliffs, NJ, 1967.

[16] H. WEBER, *On the numerical approximation of secondary bifurcation problems*, Lecture Notes in Mathematics 878, E. L. Allgower, K. Glasshoff and H. O. Peitgen, eds., Springer-Verlag, Berlin, Heidelberg, New York, 1980.

SIAM J. MATH. ANAL.
Vol. 17, No. 3, May 1986

006

A PRIORI BOUNDS AND EXISTENCE OF POSITIVE SOLUTIONS FOR SINGULAR NONLINEAR BOUNDARY VALUE PROBLEMS*

D. R. DUNNINGER† AND J. C. KURTZ†

Abstract. We consider singular nonlinear problems of the form $-(1/p)(pu')'=f(u)$ on $(0,1)$ with Dirichlet or mixed boundary conditions. The existence of a positive solution is obtained by first deriving a priori bounds in an appropriate weighted Sobolev norm, which in turn imply L^∞ a priori bounds. We then apply known results concerning fixed points of mappings of a cone.

1. Introduction. In this paper we seek the existence of solutions to singular problems of the form

$$Lu \equiv -\frac{1}{p}(pu')' = f(u) \quad \text{on } (0,1)$$

with $u(1)=0$ and boundary conditions at $t=0$ which will be determined by the behaviour of $p(t)$ near $t=0$. We will always assume at least that $p \in C[0,1] \cap C^1(0,1]$, $p(0)=0$ and $p(t)>0$ on $(0,1]$.

Linearly independent solutions of $Lu=0$ are $u_1(t)\equiv 1$ and $u_2(t)=\int_t^1 (ds/p(s))$. Thus, if p satisfies $\int_0^1 (ds/p(s))<\infty$, then u_1, u_2 are bounded and we may consider the Dirichlet problem

$$\begin{aligned} &Lu=f(u) \quad \text{on } (0,1),\\ &u(0)=u(1)=0. \end{aligned} \tag{1.1}$$

On the other hand, if $\int_0^1 (ds/p(s))=\infty$, then u_2 becomes unbounded near $t=0$, and the Dirichlet problem is inappropriate. In this case we take as our model the Dirichlet problem in $R^n (n \geqq 3)$

$$\begin{aligned} &-\Delta u = f(u), \qquad |x|<1,\\ &u=0 \quad \text{on } |x|=1. \end{aligned}$$

The equation for radial solutions $u=u(r)$, $r=|x|$, is then

$$\begin{aligned} &-r^{1-n}(r^{n-1}u')' = f(u), \qquad r \in (0,1),\\ &u(1)=0 \end{aligned}$$

and one seeks solutions satisfying $u'(0)=0$ so that the solution $u=u(|x|)$ is C^2 at the origin. Thus we are led to consider the problem

$$\begin{aligned} &Lu=f(u) \quad \text{on } (0,1),\\ &u(1)=0, \qquad \lim_{t\to 0^+} p(t)u'(t)=0. \end{aligned} \tag{1.2}$$

This weaker form of the boundary condition at $t=0$ seems more natural in the weak formulation of the problem, and will eventually lead to $u'(0)=0$.

*Received by the editors August 25, 1983, and in revised form October 15, 1984.
† Department of Mathematics, Michigan State University, East Lansing, Michigan 48824.

In the sequel it will become clear that problem (1.2) is also appropriate in the case $\int_0^1 (dt/p(t)) < \infty$. Thus we will consider (1.2) in both cases.

We shall denote the problems (1.1) and (1.2) collectively by

$$\begin{aligned} &Lu = f(u) \quad \text{on } (0,1), \\ &u \in \mathscr{B}, \end{aligned} \tag{1.3}$$

where $u \in \mathscr{B}$ means that u satisfies the appropriate boundary conditions depending on the behaviour of p near $t=0$.

For regular boundary value problems, fairly general conditions on the function f which ensure the existence of a solution have recently been given in [3], [4]. This approach is being pursued by the authors in a separate paper [1] for the singular problem.

In the present paper we consider a much different class of nonlinearities, and prove the existence of a positive solution by first deriving a priori bounds for any positive solution. Our approach is generally along the lines of [2] and [11]. Existence alone can be obtained under weaker hypotheses (see [2, Thm. 2.3]).

Sections 2 and 3 contain definitions and a number of technical results, which will be needed later, concerning the operator L and the various spaces involved. In §4 we obtain a priori L^∞-bounds for positive solutions of (1.3). Finally, in §5 we use the a priori bounds to obtain the existence of a positive solution by a fixed point argument.

2. Definitions and imbedding theorems. For $i=1,2$, we define the spaces $(X_i, \|\cdot\|_X)$ by

$$X_i = \{ u \mid u \in C[0,1],\ pu' \in C[0,1],\ Lu \in C[0,1],\ u \in \mathscr{B}_i \}$$

where

$$\begin{aligned} \mathscr{B}_1 &= \Big\{ u \Big| u(1) = \lim_{t \to 0^+} p(t)u'(t) = 0 \Big\}, \\ \mathscr{B}_2 &= \{ u \mid u(0) = u(1) = 0 \}, \end{aligned}$$

and

$$\|u\|_X = |u|_0 + |pu'|_0 + |Lu|_0,$$

where as usual

$$|u|_0 = \max_{0 \leqq t \leqq 1} |u(t)|.$$

Correspondingly, we define $(Y_i, \|\cdot\|_Y)$ by

$$\begin{aligned} Y_1 &= \{ u \mid u \in AC_{\text{loc}}(0,1],\ u(1) = 0,\ \|u\|_Y < \infty \}, \\ Y_2 &= \{ u \mid u \in AC_{\text{loc}}(0,1],\ u \in \mathscr{B}_2,\ \|u\|_Y < \infty \}, \end{aligned}$$

where

$$\|u\|_Y = \left(\int_0^1 u'(t)^2 p(t)\, dt \right)^{1/2}.$$

We denote by L_p^q the space of all measurable functions u for which

$$\|u\|_{L_p^q}=\left(\int_0^1|u(t)|^q p(t)\,dt\right)^{1/q}<\infty$$

and by $W_p^{1,2}$ the space of functions $u\in AC_{\text{loc}}(0,1]$ for which $u,u'\in L_p^2$ with

$$\|u\|_{W_p^{1,2}}=\left\{\|u\|_{L_p^2}^2+\|u'\|_{L_p^2}^2\right\}^{1/2}.$$

Remark 2.1. If we assume the condition

$$\int_0^1\frac{1}{p(s)}\int_0^t p(t)\,dt\,ds<\infty, \tag{P1}$$

then $\|u\|_Y$ is equivalent to $\|u\|_{W_p^{1,2}}$. In fact, for $u\in Y_i$ we have

$$|u(t)|\leqq\left(\int_t^1 u'(\tau)^2 p(\tau)\,d\tau\right)^{1/2}\left(\int_t^1\frac{d\tau}{p(\tau)}\right)^{1/2}, \tag{2.1}$$

and thus

$$\int_0^1 u(t)^2 p(t)\,dt\leqq\|u\|_Y^2\int_0^1\frac{1}{p(\tau)}\int_0^\tau p(t)\,dt\,d\tau.$$

We now state and prove a number of imbedding theorems.

LEMMA 2.1. *If* (P1) *holds, the imbedding* $i: X_1\to C[0,1]$ *is compact.*

Proof. For $u\in X_1$ we have

$$|p(t)u'(t)|\leqq\int_0^t|Lu|p(\tau)\,d\tau\leqq\|u\|_X\int_0^t p(\tau)\,d\tau \tag{2.2}$$

from which it follows that

$$|u(t_2)-u(t_1)|\leqq\|u\|_X\int_{t_1}^{t_2}\frac{1}{p(t)}\int_0^t p(\tau)\,d\tau\,dt$$

and

$$|u(t)|\leqq\|u\|_X\int_0^1\frac{1}{p(t)}\int_0^t p(\tau)\,d\tau\,dt.$$

The result now follows from the Arzela–Ascoli theorem.

LEMMA 2.2. *If* p *satisfies*

$$\int_0^1\frac{dt}{p(t)}<\infty \tag{P2}$$

the imbedding $i: X_2\to C[0,1]$ *is compact.*

Proof. If $u\in X_2$ we have

$$|u(t_2)-u(t_1)|\leqq\|u\|_X\int_{t_1}^{t_2}\frac{dt}{p(t)}$$

and

$$|u(t)|\leqq\|u\|_X\int_0^1\frac{dt}{p(t)}$$

and the result follows again from the Arzela–Ascoli theorem.

LEMMA 2.3. *If* (P1) ((P2)) *holds, then the imbedding* $i: X_1 \to Y_1$ ($i: X_2 \to Y_2$) *is continuous.*

Proof. In the first case we have from (2.2)

$$\int_0^1 u'(t)^2 p(t)\,dt \leqq \|u\|_X^2 \int_0^1 \frac{1}{p(t)} \left(\int_0^t p(\tau)\,d\tau \right)^2 dt$$

$$\leqq C\|u\|_X^2 \int_0^1 \frac{1}{p(t)} \int_0^t p(\tau)\,d\tau\,dt$$

while in the second case we have

$$\int_0^1 u'(t)^2 p(t)\,dt = \int_0^1 \frac{1}{p(t)} \left(p(t)u'(t) \right)^2 dt \leqq \|u\|_X^2 \int_0^1 \frac{dt}{p(t)}.$$

LEMMA 2.4. *If* (P2) *holds, the imbedding* $i: Y_2 \to C[0,1]$ *is compact.*

Proof. For $u \in Y_2$ we have from (2.1)

$$|u(t)| \leqq \|u\|_Y \left(\int_0^1 \frac{dt}{p(t)} \right)^{1/2}.$$

Since

$$|u(t_2) - u(t_1)| \leqq \|u\|_Y \left(\int_{t_1}^{t_2} \frac{dt}{p(t)} \right)^{1/2}$$

the result follows once more from the Arzela–Ascoli theorem.

In order to prove a corresponding imbedding theorem for Y_1 we must impose more restrictions on $p(t)$, namely

(P3) $$\lim_{t \to 0^+} \frac{tp'(t)}{p(t)} = \alpha > 0,$$

(P4) $$C_1 \left(1 + |\log t| \right)^{-1} \leqq \frac{t^\alpha}{p(t)} \leqq C_2 \left(1 + |\log t| \right).$$

Remark 2.2. It follows from (P3) that

$$\lim_{t \to 0^+} \frac{1}{p(t)} \int_0^t p(\tau)\,d\tau = \lim_{t \to 0^+} \frac{p(t)}{p'(t)} = 0$$

so that (P1) also holds. In addition, it follows easily from (P4) that (P2) holds for $0 < \alpha < 1$, whereas if $\alpha \geqq 1$,

$$\int_0^1 \frac{ds}{p(s)} = \infty.$$

We now define

$$\sigma + 1 = \sup \left\{ q \,\middle|\, \int_0^1 \left(\int_t^1 \frac{d\tau}{p(\tau)} \right)^{q/2} p(t)\,dt < \infty \right\}.$$

LEMMA 2.5. *Let* (P3), (P4) *hold. Then*

$$\sigma + 1 = \begin{cases} +\infty & \text{if } 0 < \alpha \leqq 1, \\ \dfrac{2(\alpha+1)}{\alpha - 1} & \text{if } \alpha > 1. \end{cases}$$

Proof. For $0<\alpha<1$, (P2) holds, and therefore $\sigma+1=\infty$.

For $\alpha=1$,

$$\int_t^1 \frac{d\tau}{p(t)} \leqq C_2 \int_t^1 \frac{(1-\log\tau)}{\tau} d\tau \leqq C(1-\log t)^2$$

and therefore,

$$\int_0^1 \left(\int_t^1 \frac{d\tau}{p(\tau)} \right)^{q/2} p(t)\, dt \leqq C \int_0^1 t(1-\log t)^{1+q} dt < \infty$$

for every q, which implies $\sigma+1=+\infty$.

For $\alpha>1$,

$$\int_t^1 \frac{d\tau}{p(\tau)} \leqq C_2 \int_t^1 \tau^{-\alpha}(1-\log\tau)\, d\tau \leqq C t^{1-\alpha}(1-\log t)$$

and therefore,

$$\int_0^1 \left(\int_t^1 \frac{d\tau}{p(t)} \right)^{q/2} p(t)\, dt \leqq C \int_0^1 t^{\alpha+(1-\alpha)q/2}(1-\log t)^{1+(q/2)} dt.$$

The last integral, however, is finite if $\alpha+(1-\alpha)q/2>-1$, i.e., $q<2(\alpha+1)/(\alpha-1)$. Thus $\sigma+1 \geqq 2(\alpha+1)/(\alpha-1)$. Similarly we can show that $\sigma+1 \leqq 2(\alpha+1)/(\alpha-1)$ which completes the proof.

LEMMA 2.6. *If* (P3), (P4) *hold, then the imbedding* $i: Y_1 \to L_p^q$ *is compact for* $2 \leqq q < \sigma+1$.

Proof. Using (2.1) we obtain

$$\int_0^\varepsilon |u(t)|^q p(t)\, dt \leqq \|u\|_Y^q \int_0^\varepsilon p(t) \left(\int_t^1 \frac{d\tau}{p(\tau)} \right)^{q/2} dt = o(1)\|u\|_Y^q \quad \text{as } \varepsilon \to 0^+. \tag{2.3}$$

Assuming $\|u_n\|_Y \leqq C$, it follows from Remark 2.1 that $\|u_n\|_{W_p^{1,2}}$ is also uniformly bounded. Since for every $\varepsilon>0$, $W_p^{1,2} \subseteq W^{1,2}([\varepsilon,1])$, the usual Sobolev space, it follows that there exists a subsequence, which we again call $\{u_n\}$, and a function $u \in Y_1$ such that $u_n \to u$ uniformly on compact subsets of $(0,1]$. From (2.3) we can easily infer that $\|u_n-u\|_Y \to 0$, which completes the proof.

LEMMA 2.7. *Let* (P3) *hold and, in addition, assume*

$$C_1(1+|\log t|)^{-1} \leqq \frac{t^\alpha}{p(t)} \leqq C_2, \tag{P4}'$$

$$p'(t)>0 \quad \text{on } (0,1). \tag{P5}$$

Then, if $\alpha>1$, *the imbedding* $i: Y_1 \to L_p^{\sigma+1}$ *is continuous.*

Proof. For $u \in Y_1$, set $v=|u|^{2\alpha/(\alpha-1)}$. Then

$$\begin{aligned} |v(t)| &\leqq \frac{1}{p(t)} \int_t^1 |v'(\tau)|\, p(\tau)\, d\tau, \\ |v(t)|^{1+(1/\alpha)} &\leqq C \int_t^1 |v(\tau)|^{1/\alpha} |v'(\tau)|\, d\tau, \end{aligned} \tag{2.4}$$

and therefore,

$$\int_0^1 |v(t)|^{1+(1/\alpha)} p(t)\,dt \leqq C\int_0^1 p(t)\int_t^1 |v(\tau)|^{1/\alpha}|v'(\tau)|\,d\tau\,dt$$

$$\leqq C\sup_{\tau>0}\left\{|v(\tau)|^{1/\alpha}\frac{1}{p(\tau)}\int_0^\tau p(t)\,dt\right\}\int_0^1 |v'(\tau)|p(\tau)\,d\tau.$$

Using (2.4) and (P4)′ we obtain

$$\int_0^1 |v(t)|^{1+(1/\alpha)} p(t)\,dt \leqq C\left(\int_0^1 |v'(\tau)|p(\tau)\,d\tau\right)^{1+(1/\alpha)} \sup_{\tau>0}\left\{p(\tau)^{-1-(1/\alpha)}\int_0^\tau p(t)\,dt\right\}$$

$$\leqq C\left(\int_0^1 |v'(\tau)|p(\tau)\,d\tau\right)^{1+(1/\alpha)}.$$

Now

$$|v'(t)| \leqq \frac{2\alpha}{\alpha-1}|u(t)|^{(\alpha+1)/(\alpha-1)}|u'(t)|.$$

Thus

$$\left(\int_0^1 |u(t)|^{\sigma+1} p(t)\,dt\right)^{\alpha/(\alpha+1)} \leqq C\int_0^1 |u(t)|^{(\alpha+1)/(\alpha-1)}|u'(t)|p(t)\,dt$$

$$\leqq C\left(\int_0^1 u'(t)^2 p(t)\,dt\right)^{1/2}\left(\int_0^1 |u(t)|^{\sigma+1} p(t)\,dt\right)^{1/2}$$

and finally

$$\left(\int_0^1 |u(t)|^{\sigma+1} p(t)\,dt\right)^{1/(\sigma+1)} \leqq C\left(\int_0^1 u'(t)^2 p(t)\,dt\right)^{1/2}.$$

3. Further preliminary results. In this section we examine some properties of the operator L and the solutions of problem (1.3).

For $i=1,2$, let L_i be the operator given by $L_i \equiv L: X_i \to C[0,1]$. Then we have the following:

LEMMA 3.1. *If* (P1) ((P2)) *holds, then* L_1 (L_2) *is bounded, one-to-one, onto, and has an inverse* G_1 (G_2) *which is compact as an operator on* $C[0,1]$. *Moreover,*

$$G_i h(t) = \int_0^1 g_i(t,s)h(s)p(s)\,ds, \qquad h \in C[0,1],$$

where

$$g_1(t,s) = \begin{cases} \displaystyle\int_s^1 \frac{d\tau}{p(\tau)}, & t \leqq s, \\ \displaystyle\int_t^1 \frac{d\tau}{p(\tau)}, & t \geqq s, \end{cases}$$

$$g_2(t,s) = \begin{cases} \displaystyle\left(\int_0^1 \frac{d\tau}{p(\tau)}\right)^{-1}\int_s^1 \frac{d\tau}{p(\tau)}\int_0^t \frac{d\tau}{p(\tau)}, & t \leqq s, \\ \displaystyle\left(\int_0^1 \frac{d\tau}{p(\tau)}\right)^{-1}\int_t^1 \frac{d\tau}{p(\tau)}\int_0^s \frac{d\tau}{p(\tau)}, & t \geqq s. \end{cases}$$

Proof. Since $|L_i u|_0 \leqq \|u\|_X$, it follows that each L_i is bounded and one-to-one. By a direct computation we find that $G_i: C[0,1] \to X_i$ and $G_i = L_i^{-1}$. The compactness of the G_i follows from the closed graph theorem and Lemmas 2.1 and 2.2.

LEMMA 3.2. *Let* $f \in C(\mathbb{R})$. (i) *If* $u \in X_1$ *is a positive solution of* (1.2), *then* $u'(t) < 0$ *on* $(0,1]$. (ii) *If* $u \in X_2$ *is a positive solution of* (1.1), *then* $u'(t) > 0$ *on* $[0, t_0)$ *and* $u'(t) < 0$ *on* $(t_0, 1]$ *for some* $t_0 \in (0,1)$.

Proof. The proofs follow along the lines of [7, Lemma 4.1].

LEMMA 3.3. *Let* (P3), (P4) *hold. Suppose* $f \in C(\mathbb{R})$ *and* $f(t) = O(t^k)$ *for some* $k \in [1, \sigma)$. *If* $u \in Y_1$ *is a weak solution of* (1.2), *then in fact* $u \in C^2[0,1]$ *and* $u'(0) = 0$.

Proof. We recall first that $u \in Y_i$ is a weak solution of (1.3) if

$$\int_0^1 u'(t) w'(t) p(t)\, dt = \int_0^1 f(u(t)) w(t) p(t)\, dt \tag{3.1}$$

holds for all $w \in Y_i$. In view of Lemma 2.4, Lemma 2.6 and the estimate $|f(u)| \leqq C_1 + C_2 |u|^k$ it follows in standard fashion that the right-hand side of (3.1) is well defined.

Next we note that in fact $u \in C^2(0,1]$ and satisfies (1.2), cf. [9, Lemma 1, p. 209].

We now show that $u \in L^\infty$. If $\alpha < 1$, then by Remark 2.2, $\int_0^1 (ds/p(s)) < \infty$ and the result follows from (2.1).

Now suppose $\alpha \geqq 1$. Integrating (1.2) twice gives

$$u(t) = \int_t^1 \frac{1}{p(\tau)} \int_0^\tau f(u(s)) p(s)\, ds\, d\tau. \tag{3.2}$$

Since $|f(u)| \leqq C_1 + C_2 |u|^k$, (3.2) yields

$$|u(t)| \leqq C_1 + C_2 \int_t^1 \frac{1}{p(\tau)} \int_0^\tau |u(s)|^k p(s)\, ds\, d\tau. \tag{3.3}$$

We now iterate (3.3), beginning with the initial estimate

$$u(t) = O\left(t^{(1-\alpha)/2} (1 - \log t)^{1/2}\right)$$

which follows from (2.1) and (P4). If at the ith stage we have

$$u(t) = O\left(t^{\lambda_i} (1 - \log t)^{\beta_i}\right),$$

where $\lambda_i \geqq 1 - \alpha/2$ and $\beta_i > 0$, then (3.3) yields (for $\alpha > 1$):

$$|u(t)| \leqq C_1 + C_2 \int_t^1 \tau^{1 + k\lambda_i} (1 - \log \tau)^{k\beta_i + 2} d\tau$$

$$= \begin{cases} O(1), & 1 + k\lambda_i > -1, \\ O\left(t^{2 + k\lambda_i} (1 - \log t)^{k\beta_i + 2}\right), & 1 + k\lambda_i < -1, \\ O\left((1 - \log t)^{k\beta_i + 3}\right), & 1 + k\lambda_i = -1. \end{cases}$$

Thus $\lambda_i \geqq (1 - \alpha)/2$ and $1 + k\lambda_i < -1$ imply $\lambda_{i+1} = k\lambda_i + 2$. Moreover, it is readily seen that $\lambda_{i+1} > \lambda_i$. If $1 + k\lambda_i < -1$ for all i, then λ_i converges to some λ and $\lambda = k\lambda + 2$. But this means

$$\lambda = \frac{2}{1-k} < \frac{2}{1-\sigma} = \frac{1-\alpha}{2},$$

which is impossible. If at some stage $1+k\lambda_i=-1$, then one more iteration gives $u\in L^\infty$. Thus after finitely many iterations $1+k\lambda_i\geqq -1$, and we have $u\in L^\infty$. The case $\alpha=1$ is similar.

For the second part of the lemma we observe first that

$$u(0)=\int_0^1 \frac{1}{p(\tau)}\int_0^\tau f(u(s))p(s)\,ds\,d\tau.$$

Hence

$$\left|\frac{u(t)-u(0)}{t}\right|\leqq \frac{C}{t}\int_0^t \frac{1}{p(\tau)}\int_0^\tau p(s)\,ds\,d\tau,$$

which by Remark 2.2 implies $u'(0)=0$. Next we see that

$$\frac{u'(t)}{t}=-\frac{1}{tp(t)}\int_0^t f(u(s))p(s)\,ds,$$

which by l'Hôpital's rule and (P3) gives

$$u''(0)=-\lim_{t\to 0^+}\frac{f(u(t))p(t)}{p(t)+tp'(t)}=-\frac{f(u(0))}{\alpha+1}.$$

On the other hand, we have from the equation

$$\lim_{t\to 0^+}u''(t)=\lim_{t\to 0^+}\left\{-f(u(t))-\left(\frac{tp'(t)}{p(t)}\right)\left(\frac{u'(t)}{t}\right)\right\}=-\frac{f(u(0))}{\alpha+1}=u''(0),$$

and therefore $u\in C^2[0,1]$.

Finally, we shall need a result concerning eigenvalues and eigenfucntions of (1.3).

First we recall that if for a complex number λ there is a nontrivial $v\in Y_i$ such that

$$a_i(v,w)\equiv\int_0^1 v'(t)w'(t)p(t)\,dt=\lambda\int_0^1 v(t)w(t)p(t)\,dt$$

holds for all $w\in Y_i$, then we call λ an eigenvalue of L and we call v a generalized eigenfunction (corresponding to λ).

Under the assumption (P1), it follows from Remark 2.1 that the bilinear from $a_i(v,w)$ is positive definite. Moreover, assuming (P2)–(P4), it follows from Lemmas 2.4 and 2.6 that the imbedding $i: Y_i\to L_p^2$ is compact. Thus by a standard result [10], there exists a sequence of eigenvalues and generalized eigenfunctions of L.

In addition, we have the well-known result:

LEMMA 3.4. *Assume* (P2)–(P4). *Let* λ_1 *be the first eigenvalue of L with corresponding generalized eigenfunction* $v_1\in Y_i$. *Then* $\lambda_1>0$ *and*

$$\lambda_1=\min_{\substack{u\in Y_i\\ u\neq 0}}\frac{\int_0^1 u'(t)^2p(t)\,dt}{\int_0^1 u(t)^2p(t)\,dt}.$$

Moreover, v_1 *may be chosen to be positive in* $(0,1)$.

Remark 3.1. It follows from Lemmas 3.2–3.4 that

(i) For problem (1.2) the eigenfunction v_1 is positive and strictly decreasing.

(ii) For problem (1.1) the eigenfunction v_1 is positive and strictly increasing on $[0,t_0]$, strictly decreasing on $[t_0,1]$, for some $t_0\in(0,1)$.

It follows also from (1.1) (1.2) that $v_1\in X_1$ ($v_1\in X_2$).

4. A priori bounds. We consider first problem (1.1).

THEOREM 4.1. *Suppose*

$$\int_0^1 \frac{dt}{p(t)} < \infty \tag{P2}$$

holds and f satisfies

$$f \in C(\mathbb{R}^+), \tag{F1}$$

$$\liminf_{t\to\infty} \frac{f(t)}{t} > \lambda_1. \tag{F2}$$

Then any positive solution $u \in X_2$ of (1.1) *satisfies $|u|_0 \leqq C$, where C is independent of u.*

Here and in all that follows, C will denote a constant, different on different appearances, which is independent of u.

Proof. Multiplying (1.1) by pv_1 and integrating by parts gives

$$\int_0^1 u'v_1' p\,dt = \int_0^1 f(u)v_1 p\,dt. \tag{4.1}$$

Similarly, multiplying $Lv_1 = \lambda_1 v_1$ by pu and integrating by parts yields

$$\int_0^1 u'v_1' p\,dt = \lambda_1 \int_0^1 uv_1 p\,dt$$

which, taken together with (4.1) gives

$$\lambda_1 \int_0^1 uv_1 p\,dt = \int_0^1 f(u)v_1 p\,dt. \tag{4.2}$$

It follows from (F2) that there exist $s_0 > 0$, $\lambda > \lambda_1$ such that $f(s) \geqq \lambda s$ whenever $s \geqq s_0$. Thus (4.2) yields

$$\lambda \int_0^1 uv_1 p\,dt \leqq \int_0^1 f(u)v_1 p\,dt + C = \lambda_1 \int_0^1 uv_1 p\,dt + C$$

so that

$$\int_0^1 uv_1 p\,dt \leqq \frac{C}{\lambda - \lambda_1}, \qquad \int_0^1 f(u)v_1 p\,dt \leqq \frac{C\lambda_1}{\lambda - \lambda_1},$$

and finally also

$$\int_0^1 |f(u)|v_1 p\,dt \leqq C. \tag{4.3}$$

By Remark 3.1 and Lemma 3.2 we have for $0 \leqq t \leqq \varepsilon$, ε sufficiently small,

$$C \geqq \int_0^1 uv_1 p\,d\tau \geqq \int_\varepsilon^{2\varepsilon} uv_1 p\,d\tau \geqq u(t)\varepsilon v_1(\varepsilon) \min_{\varepsilon \leqq \tau \leqq 2\varepsilon} p(\tau)$$

so that $u(t) \leqq C$ for $0 \leqq t \leqq \varepsilon$. Similarly, $u(t) \leqq C$ for $1 - \varepsilon \leqq t \leqq 1$. It follows easily from (4.3) that

$$\int_0^1 |f(u)|p\,dt \leqq C.$$

Next, since $u = G_1 f(u)$ by Lemma 3.1,

$$p(t)u'(t) = \left(\int_0^1 \frac{d\tau}{p(\tau)}\right)^{-1} \left\{ \int_t^1 \left(\int_s^1 \frac{d\tau}{p(\tau)} \right) f(u(s)) p(s)\, ds \right.$$
$$\left. - \int_0^t \left(\int_0^s \frac{d\tau}{p(\tau)} \right) f(u(s)) p(s)\, ds \right\}$$

which in view of (P2) yields

$$|p(t)u'(t)| \leqq \int_0^1 |f(u(s))|\, p(s)\, ds \leqq C.$$

Thus

$$|u(t)| = \left| - \int_t^1 \frac{1}{p(\tau)} (p(\tau)u'(\tau))\, d\tau \right|$$
$$\leqq \left(\int_0^1 |f(u(\tau))|\, p(\tau)\, d\tau \right) \left(\int_0^1 \frac{d\tau}{p(\tau)} \right) \leqq C$$

and the theorem is proved.

Next we turn our attention to problem (1.2).

THEOREM 4.2. *Suppose*

(P3) $$\lim_{t \to 0^+} \frac{tp'(t)}{p(t)} = \alpha > 0,$$

(P4)′ $$C_1 (1 + |\log t|)^{-1} \leqq \frac{t^\alpha}{p(t)} \leqq C_2,$$

(P5) $$p'(t) > 0 \quad \text{on } (0,1),$$

hold and f satisfies (F1), (F2) *and*

(F3) $$\lim_{t \to \infty} t^{-\sigma} f(t) = 0 \qquad (\alpha > 1),$$
$$f(t) = O(t^r) \quad \text{for some } r \in (0, \infty), \text{ as } t \to \infty, \ (0 < \alpha \leqq 1),$$

(F4) $$\overline{\lim_{t \to \infty}} \frac{tf(t) - \theta F(t)}{t^2 f(t)^{2/(\alpha+1)}} \leqq 0,$$
$$\text{for some } \theta \in [0, \sigma + 1) \text{ if } \alpha > 1, \text{ where } F(t) = \int_0^t f(s)\, ds.$$

Then any positive solution $u \in X_1$ of (1.2) *satisfies* $|u|_0 \leqq C$.

Remark 4.1. Consider the typical case $p(t) = t^{n-1}$ $(n \geqq 3)$, where $\alpha = n-1$, $\sigma = (n+2)/(n-2)$. Since (F4) is satisfied in the case where $t^{-\theta} F(t)$ is decreasing for large t, it follows easily that (F1)–(F4) are satisfied for $f(u) = u^k$, $1 < k < (n+2)/(n-2)$.

Proof of Theorem 4.2. The proof is very similar to the proof of [2, Thm. 1.1]. We include the details for the sake of completeness.

As in the proof of Theorem 4.1 we have

(4.4) $$\int_0^1 |f(u)|\, p\, dt \leqq C.$$

Multiplying (1.2) by p and integrating gives

$$p(t)u'(t) = -\int_t^1 f(u(\tau))p(\tau)\,d\tau$$

and thus

$$|u'(t)| \leqq \frac{1}{p(1-\varepsilon)}\int_0^1 |f(u)|p\,d\tau \leqq C_\varepsilon$$

for $1-\varepsilon \leqq t \leqq 1$.

We now consider the case $\alpha > 1$ and obtain a "Pohozaev" type identity. Multiplying (1.2) by u' and integrating we get

$$\frac{1}{2}\{u'(1)^2 - u'(t)^2\}p(t) + p(t)\int_t^1 \frac{p'(\tau)}{p(\tau)}u'(\tau)^2 d\tau,$$

which implies

$$\begin{aligned}\int_0^1 F(u(t))p(t)\,dt &= \frac{1}{2}u'(t)^2\int_0^1 p(t)\,dt - \frac{1}{2}\int_0^1 u'(t)^2 p(t)\,dt \\ &\quad + \int_0^1 p(t)\int_t^1 \frac{p'(\tau)}{p(\tau)}u'(\tau)^2 d\tau\,dt \\ &= \frac{1}{2}u'(1)^2 P(1) + \int_0^1 \left\{\frac{P(t)p'(t)}{p^2(t)} - \frac{1}{2}\right\}u'(t)^2 p(t)\,dt\end{aligned}$$

where

$$P(t) = \int_0^t p(\tau)\,d\tau.$$

It follows from (P3) and l'Hôpital's rule that

$$\lim_{t\to 0^+}\frac{P(t)p'(t)}{p^2(t)} - \frac{1}{2} = \frac{1}{\sigma+1}.$$

Thus we may choose $\eta \in (\theta, \sigma+1)$ so that

$$\frac{P(t)p'(t)}{p^2(t)} - \frac{1}{2} \leqq \frac{1}{\eta} \quad \text{on } (0,\varepsilon),$$

and

$$\int_0^1 F(u)\,p\,dt \leqq C_\varepsilon + \frac{1}{\eta}\int_0^1 (u')^2 p\,dt \leqq C_\varepsilon + \frac{1}{\eta}\int_0^1 uf(u)\,p\,dt. \tag{4.5}$$

Using (F4) we have for every $\varepsilon > 0$ the existence of a $t_\varepsilon > 0$ such that

$$tf(t) \leqq \theta F(t) + \varepsilon t^2 |f(t)|^{2/(\alpha+1)} \quad \text{for } t \geqq t_\varepsilon.$$

This together with (4.5) gives

$$\begin{aligned}\int_0^1 f(u)up\,dt &\leqq \theta\int_0^1 F(u)\,p\,dt + \varepsilon\int_0^1 u^2 |f(u)|^{2/(\alpha+1)} p\,dt + C_\varepsilon \\ &\leqq \frac{\theta}{\eta}\int_0^1 f(u)up\,dt + \varepsilon\int_0^1 u^2 |f(u)|^{2/(\alpha+1)} p\,dt + C_\varepsilon.\end{aligned} \tag{4.6}$$

But, by (4.4) and Lemma 2.7, we have

$$\int_0^1 u^2 |f(u)|^{2/(\alpha+1)} p\,dt \leqq \left(\int_0^1 |f(u)|p\,dt\right)^{2/(\alpha+1)} \left(\int_0^1 |u|^{\sigma+1} p\,dt\right)^{(\alpha-1)/(\alpha+1)}$$
$$\leqq C\int_0^1 (u')^2 p\,dt = C\int_0^1 f(u)up\,dt.$$

Combining this with (4.6) gives

$$\int_0^1 f(u)up\,dt \leqq \frac{\theta}{\eta}\int_0^1 f(u)up\,dt + C\varepsilon\int_0^1 f(u)up\,dt + C_\varepsilon$$

and thus

$$\left(1-\frac{\theta}{\eta}-C\varepsilon\right)\int_0^1 f(u)up\,dt \leqq C_\varepsilon.$$

Choosing $\varepsilon > 0$ sufficiently small gives

$$\int_0^1 f(u)up\,dt = \int_0^1 (u')^2 p\,dt \leqq C.$$

Next we observe that for $r > 1$

$$r\int_0^1 (u')^2 u^{r-1} p\,dt = \frac{4r}{(r+1)^2}\int_0^1 \left|(u^{(r+1)/2})'\right|^2 p\,dt = \int_0^1 f(u)u^r p\,dt. \tag{4.7}$$

According to (F3) we have

$$f(u)u^r \leqq \varepsilon u^{\sigma+r} + C_\varepsilon. \tag{4.8}$$

Setting $q = (\sigma+1)((r-1)/2)$ we obtain from (4.7), (4.8) and Lemma 2.7

$$\left(\int_0^1 u^q p\,dt\right)^{2/(\sigma+1)} = \left(\int_0^1 (u^{(r+1)/2})^{\sigma+1} p\,dt\right)^{2/(\sigma+1)}$$
$$\leqq C\int_0^1 \left|(u^{(r+1)/2})'\right|^2 p\,dt = C\int_0^1 f(u)u^r p\,dt$$
$$\leqq C\varepsilon\int_0^1 u^{r+1}u^{\sigma-1} p\,dt + C_\varepsilon$$
$$\leqq C\varepsilon\left(\int_0^1 u^q p\,dt\right)^{2/(\sigma+1)} \left(\int_0^1 u^{\sigma+1} p\,dt\right)^{(\sigma-1)/(\sigma+1)} + C_\varepsilon$$
$$\leqq C\varepsilon\left(\int_0^1 u^q p\,dt\right)^{2/(\sigma+1)} \left(\int_0^1 (u')^2 p\,dt\right)^{(\sigma-1)/2} + C_\varepsilon$$
$$\leqq C\varepsilon\left(\int_0^1 u^q p\,dt\right)^{2/(\sigma+1)} + C_\varepsilon.$$

It follows that

$$\left(\int_0^1 u^q p\,dt\right)^{1/q} \leqq C \tag{4.9}$$

for $q=(\sigma+1)((r+1)/2)$, for all $r>1$; i.e. for all $q>\sigma+1$. Hence (4.9) is valid for all $q\geqq 1$. Finally, we have from (3.1) and (4.9)

$$\begin{aligned}|u(t)| &\leqq \int_0^1 \frac{1}{p(t)} \int_0^t |f(u(s))|p(s)\,ds\,dt \\ &= \int_0^1 |f(u(s))| \left(\int_s^1 \frac{dt}{p(t)} \right) p(s)\,ds \\ &\leqq \left(\int_0^1 |f(u(s))|^{r/(r-2)} p(s)\,ds \right)^{1-(2/r)} \left(\int_0^1 p(s) \left(\int_s^1 \frac{dt}{p(t)} \right)^{r/2} \right)^{2/r}.\end{aligned}$$

Taking $2<r<\sigma+1$ gives

$$|u(t)| \leqq C_1 \left(\int_0^1 u^{r\sigma/(r-2)} p\,ds \right)^{1-(2/r)} + C_2 \leqq C.$$

According to Lemma 2.5, $\sigma=+\infty$ when $\alpha=1$. Thus in view of (F3) and Lemma 2.6 it follows as above that $|u|_0 \leqq C$.

For the case $0<\alpha<1$, according to (F3), we may choose $\gamma\in(0,1)$ so that

$$\lim_{t\to\infty} \frac{tf(t)}{t^2 f(t)^\gamma} = 0.$$

Then for every $\varepsilon>0$ there exists a $t_\varepsilon>0$ such that

$$tf(t) \leqq \varepsilon t^2 f(t)^\gamma \quad \text{for } t\geqq t_\varepsilon.$$

Thus

$$\begin{aligned}\int_0^1 (u')^2 p\,dt = \int_0^1 f(u) up\,dt &\leqq \varepsilon \int_0^1 u^2 |f(u)|^\gamma p\,dt + C_\varepsilon \\ &\leqq \varepsilon \left(\int_0^1 |f(u)|p\,dt \right)^\gamma \left(\int_0^1 u^{2/(1-\gamma)} p\,dt \right)^{1-\gamma} + C_\varepsilon \\ &\leqq C\varepsilon \int_0^1 (u')^2 p\,dt + C_\varepsilon\end{aligned}$$

by (4.4) and Lemma 2.6, and again

$$\int_0^1 (u')^2 p\,dt \leqq C.$$

Using (3.1) again we have

$$|u(t)| \leqq \left(\int_0^1 \frac{dt}{p(t)} \right) \int_0^1 |f(u)|p\,dt \leqq C$$

in view of (4.4) and Remark 2.2. Thus the theorem is proved.

Remark 4.2. If in Theorem 4.2 we replace conditions (F3) and (F4) (in the case $\alpha>1$) by

$$(\text{F3})' \qquad \lim_{t\to\infty} t^{-l} f(t) = 0, \qquad 1<l<\sigma,$$

$$(\text{F4})' \qquad \overline{\lim_{t\to\infty}} \frac{tf(t)-\theta F(t)}{t^2 f(t)^{2/(\beta+1)}}, \qquad \theta\in[0,l+1),$$

where $\beta=(l+3)/(l-1)$, then condition (P4)′ may be replaced by the weaker (P4). The proof is essentially unchanged if we replace σ by l, α by β, and use Lemma 2.6 in place of Lemma 2.7.

5. Existence of a positive solution. We first observe that, by Lemma 3.1 and the classical maximum principle, finding a positive solution of (1.3) is equivalent to finding a nonzero solution $u\in K=\{u\mid u\in C[0,1],\ u\geqq 0\}$ of the equation

$$u=\Phi_i(u)\equiv G_i f(u). \tag{5.1}$$

Our tool in this regard is

PROPOSITION 5.1 (see [2]). *Let K be a cone in a Banach space B and $\Phi: K\to K$ a compact map with $\Phi(0)=0$. Assume that there exist numbers $0<r<R$ and a vector $v\in K-\{0\}$ such that* (i) $x\neq t\Phi(x)$ *for* $0\leqq t\leqq 1$ *and* $\|x\|=r$ *and* (ii) $x\neq\Phi(x)+tv$ *for* $t\geqq 0$ *and* $\|x\|=R$. *If* $U=\{x\in K\mid r<\|x\|<R\}$ *and* $B_\rho=\{x\in K\mid\|x\|<\rho\}$, *then Φ has a fixed point in U.*

Remark 5.1. (i) is satisfied if there exists a bounded linear map $A: B\to B$ such that $A(K)\subseteq K$, A has spectral radius less than 1, and $\Phi(x)\leqq A(x)$ for $x\in K$ and $\|x\|=r$. In addition, (ii) may be replaced by (ii)′ there exists a compact map $F:\bar{B}_R\times[0,\infty)\to K$ such that $F(x,0)=\Phi(x)$ for $\|x\|=R$, $F(x,t)\neq x$ for $\|x\|=R$ and $0\leqq t<\infty$, and $F(x,t)=x$ has no solution $x\in\bar{B}_R$ for $t\geqq t_0$.

THEOREM 5.1. *Let $f:\mathbb{R}^+\to\mathbb{R}^+$ be locally Lipschitzian and suppose the hypotheses of either Theorem* 4.1 *or Theorem* 4.2 *are satisfied. Assume in addition that*

$$\overline{\lim_{t\to 0^+}}\frac{f(t)}{t}<\lambda_1,\qquad f(0)=0. \tag{F5}$$

Then there exists at least one positive solution $u\in X_i$ of (1.3).

Proof. We seek a fixed point $u\in K$ of (5.1). By (F5) we may choose $r>0$, $\alpha<\lambda_1$ so that $f(t)\leqq\alpha t$ for $0\leqq t\leqq r$. Then $\Phi_i(u)\leqq\alpha G_i(u)$ for $|u|_0\leqq r$, and $\|\alpha G_i\|\leqq\alpha/\lambda_1<1$ so that (i) is satisfied. Next we set

$$F_i(u,t)=G_i(f(u)+t).$$

It can be shown using the techniques of the proof of Theorem 4.1 that $F_i(u,t)=u$ has no positive solutions for $t\geqq t_0$. It follows then as in §4 that the a priori bounds for positive solutions hold uniformly in t. Thus there exists $R>r$ such that $F(u,t)=u$ has no solution for $|u|_0\geqq R$. It follows from Proposition 5.1 that Φ has a fixed point $u\in U$.

REFERENCES

[1] D. R. DUNNINGER AND J. C. KURTZ, *Existence of solutions for some nonlinear singular boundary value problems*, in preparation.

[2] D. G. DE FIGUEIREDO, P. L. LIONS AND R. D. NUSSBAUM, *A priori estimates and existence of positive solutions of semilinear elliptic equations*, J. Math. Pures Appl., 61 (1982), pp. 41–63.

[3] A. GRANAS, R. B. GUEUTHER AND J. W. LEE, *Nonlinear boundary value problems for some classes of ordinary differential equations*, Rocky Mountain J. Math., 10 (1980), pp. 35–58.

[4] _____, *On a theorem of S. Bernstein*, Pacific J. Math., 74, 1 (1978), pp. 67–82.

[5] M. A. KRASNOSEL'SKII, *Fixed points of cone-compressing or cone-extending operators*, Soviet Math. Dokl., 81 (1960), pp. 1285–1288.

[6] _____, *Positive Solutions of Operator Equations*, Noordhoff, Groningen, the Netherlands, 1964.

[7] J. C. KURTZ, *Weighted Sobolev spaces with applications to singular nonlinear boundary value problems*, J. Differential Equations, 49 (1983), pp. 105–123.
[8] P. L. LIONS, *A priori estimates and existence results for some semilinear elliptic problems*, MRC Technical Summary Report, 1975, Madison, WI, 1979.
[9] V. P. MIKHAILOV, *Partial Differential Equations*, MIR Publishers, Moscow, 1978.
[10] S. G. MIKHLIN, *Mathematical Physics, an Advanced Course*, North-Holland, Amsterdam, 1970.
[11] R. D. NUSSBAUM, *Positive solutions of nonlinear elliptic boundary value problems*, J. Math. Anal. Appl., 51 (1975), pp. 461–482.
[12] R. E. L. TURNER, *A priori bounds for positive solutions of nonlinear elliptic equations in two variables*, Duke Math. J., 41 (1974), pp. 759–774.

SIAM J. MATH. ANAL.
Vol. 17, No. 3, May 1986

007

A FREE BOUNDARY PROBLEM ARISING FROM THE LEACHING OF SALINE SOILS*

PETER KNABNER†

Abstract. A multidimensional parabolic free boundary problem of the implicit type arises in models in the soil sciences. Existence, uniqueness and asymptotic behaviour of weak and classical solutions are considered.

AMS(MOS) subject classifications. Primary 35R35; secondary 35K20, 35K55, 35K60, 35K65

Key words. free boundary problems, parabolic boundary problems, nonlinear parabolic systems, parabolic systems, parabolic systems of degenerate type

1. Introduction. Approximately one-third of the developed agricultural lands in the arid and semi-arid regions of the world exhibit harmful salinity accumulation. This leads to considerably reduced crop yields (compare e.g. [4]). Reclamation of saline soils is done by leaching, the process whereby soil solution of high salt concentration is displaced by less concentrated solution until the accumulated solid salt is removed. A comprehensive exposition concerning saline soils may be found in [3]. In this paper we develop and investigate a model that describes the leaching process with special regard to the dynamics of the amount of the solid salt. One-dimensional and other simplified versions of this model appear in [9], [6], [7]: [9] compares numerical computations against experimental data, while in [6], [7] existence and uniqueness are investigated. Mathematically our multidimensional model generalizes most of the previous results in [6], [7]. Especially we have a stronger notion of solution, while we require weaker assumptions on the smoothness of the data.

2. The problem. Soil is a porous system, consisting of the solid soil matrix, the soil solution, i.e. water with dissolved substances, and the soil atmosphere, which both fill the pore space. We start with the macroscopic mass balance law for a solute, which is derived from the microscopic law by averaging over the water filled part of a "Representative Elementary Volume (REV)" (compare [1, pp. 513–520]):

$$\frac{\partial(\theta C+S)}{\partial t}=-\operatorname{div}[\mathbf{J}_1+\mathbf{J}_2+\mathbf{J}_3]+\tilde{F}, \tag{2.1}$$

where: θ is the volumetric water content, i.e. the ratio of the volume of water to the total volume of an REV, C is the concentration of the dissolved salt and S is the concentration of solids absorbed on soil surfaces or located in dead-end pore space.

$\mathbf{J}_1$ denotes the convective flux density of the solute, thus

$$\mathbf{J}_1=C\mathbf{V}, \tag{2.2}$$

where $\mathbf{V}$ is the flow rate of the soil solution. $\mathbf{J}_2$ is the dispersive flux density, due to variations of the flow velocity at the microscopic scale, and $\mathbf{J}_3$ is the flux density as a

*Received by the editors March 6, 1984. This work was supported by CNR, Italy.
† Mathematisches Institut, Universität Augsburg, 8900 Augsburg, West Germany.

result of molecular diffusion, and finally $\tilde{F}$ is the volumetric rate of supply or loss. All quantities are to be understood as averages, C over the water-filled part of an REV, while $\mathbf{V}$, S, and $\tilde{F}$ are over the total volume.

We assume

$$\mathbf{J}_2 = -\theta \mathbf{D}_1 \operatorname{grad} C, \tag{2.3}$$

$$\mathbf{J}_3 = -\theta \mathbf{D}_2 \operatorname{grad} C \quad \text{(Fick's law)}, \tag{2.4}$$

$$S = K_1 C + K_2 \quad \text{(Lapidus/Amundson, compare [2])}. \tag{2.5}$$

$\mathbf{D}_i$ are symmetric matrices, $\mathbf{D}_1$ may depend on $\mathbf{V}$ (compare [1, pp. 232–239]), K_i are functions with $K_1 \geqq 0$, both only dependent of x.

Furthermore, assuming the incompressibility of water, we can use the macroscopic balance equation for the water volume

$$\frac{\partial \theta}{\partial t} = -\operatorname{div} \mathbf{V} \tag{2.6}$$

to simplify (2.1) to:

$$AC_t - \frac{\partial}{\partial x_i}\left(D_{ij}\frac{\partial C}{\partial x_j}\right) + V_i \frac{\partial C}{\partial x_i} = \tilde{F}, \tag{2.7}$$

denoting $A := \theta + K_1$, $\mathbf{D} := \theta(\mathbf{D}_1 + \mathbf{D}_2)$ and applying Einstein's summation convention.

Let N denote the average concentration of the solid salt. Then, under assumptions described in [9], we have, as long as $N(\mathbf{x}, t) > 0$:

$$N_t = -\gamma(C^* - C). \tag{2.8}$$

C^* is the saturation concentration and γ is a known function, e.g. of the form $\tilde{\gamma}\theta$ for a constant $\tilde{\gamma} > 0$. The dissolution of N serves as a source for C, such that

$$\tilde{F} = -N_t + F. \tag{2.9}$$

For the completion of our description we need initial conditions for C and N and boundary conditions for C. The boundary conditions are assumed to be

$$D_{ij}\frac{\partial C}{\partial x_j} n_i = \beta(C - C_*) \tag{2.10}$$

with given functions β and C_*, while $\mathbf{n}$ denotes the outward normal vector. This form includes the cases (i) $\beta = V_i n_i$ and (ii) $\beta = 0$. At the portion of the boundary where we have (i), water with a salt concentration C_* flows in or out and (2.10) expresses the continuity of the mass flux density in the normal direction. In case (ii) we have either impermeability of that part of the boundary for the water and solute flux or the situation of (i) but with $C = C_*$, i.e. continuity of the concentration across the boundary. Neglecting the influence of the solute on the water flow, we can regard θ and $\mathbf{V}$ as known functions of space and time. Due to the interface between the regions $N > 0$ and

$N \equiv 0$ there is a discontinuity, since in general N_t will not vanish there. Thus a classical description has to deal with this free boundary:

Let Ω be a domain in $\mathbb{R}^n$, $n \in \mathbb{N}$, $S := \partial\Omega$, $T > 0$, $Q_{T_1T_2} := \Omega \times (T_1, T_2)$, $Q_T := Q_{0T}$, $S_T := S \times [0, T]$. Let A, D_{ij}, V_i, F, γ, C^* be given functions on $\overline{Q}_T$, C_0, N_0 are given on $\overline{\Omega}$ and β, C_* are given on S_T respectively, $i,j = 1, \cdots, n$. We will use for the elliptic part of the differential operator

$$LC := \frac{\partial}{\partial x_i}\left(D_{ij}\frac{\partial C}{\partial x_j}\right) - V_i\frac{\partial C}{\partial x_i},$$

and, $\mathbf{n}$ being the outward normal, for the outward conormal

$$\frac{\partial C}{\partial \nu} := D_{ij}\frac{\partial C}{\partial x_j}n_i.$$

For function spaces we adopt the notion of [8]; thus we use $\|\cdot\|_{q,Q_T}$ for the norm in $L_q(Q_T)$, $\|\cdot\|^{(2l)}_{q,Q_T}$ for the norm in $W_q^{2l,l}(Q_T)$ and $|\cdot|^{(l)}_{Q_T}$ for the norm in $H^{l,l/2}(\overline{Q}_T)$. In accordance with this the maximum norm is denoted by $|\cdot|^{(0)}_{Q_T}$.

DEFINITION 2.1. A triple (C, N, Φ) is called a *classical solution* in Q_T, if:

(2.11) $N \in C(\overline{Q}_T)$, $N \geqq 0$ on $\overline{Q}_T$.
Let $\Omega_{0,T} := \{x \in \overline{\Omega} \mid \text{there exists } t \in [0,T] : N(x,t) = 0\}$;
then $\Phi \in C(\Omega_{0,T})$, $\Phi(x) \in [0,T]$ for $x \in \Omega_{0,T}$.

(2.12) $N(x,t) > 0 \Leftrightarrow \Phi(x) > t$ for $x \in \Omega_{0,T}$, $t \in [0,T]$.

(2.13) Let $Q_+ := \{(x,t) \in Q_T \mid N(x,t) > 0\}$, $Q_- := Q_T \setminus \overline{Q}_+$;
then $C \in C^{2,1}(Q_+) \cap C^{2,1}(Q_-) \cap C(\overline{Q}_T)$,
$\dfrac{\partial C}{\partial x_i} \in C(\overline{Q}_T)$, $\quad i = 1, \cdots, n$, $N_t \in C(Q_+)$.

(2.14) $AC_t - LC + N_t = F$ in Q_+,
$AC_t - LC = F$ in Q_-.

(2.15) $C(x,0) = C_0(x)$, $\quad x \in \overline{\Omega}$,
$\dfrac{\partial C}{\partial \nu} = \beta(C - C_*)$ on S_T.

(2.16) $N_t = -\gamma(C^* - C)$ in Q_+.

(2.17) $N(x,0) = N_0(x)$, $\quad x \in \overline{\Omega}$.

The free boundary $\Phi(x) = t$ separates the regions $N > 0$ and $N \equiv 0$. The free boundary conditions are contained in (2.11), (2.13), assuming the continuity of N, C and $\partial C/\partial x_i$ across the interface; in particular we have $N(x, \Phi(x)) = 0$. To preserve this condition, we must have the restricted domain of definition for Φ. As we are mainly interested in the existence of classical solutions, we consider only smooth data. In the whole paper

we use

Assumption 2.2.

(2.18) $A(x,t)>0$ for all $(x,t)\in\overline{Q}_T$.

(2.19) $(D_{ij}(x,t))_{i,j}$ is symmetric and positive definite for all $(x,t)\in\overline{Q}_T$.

(2.20) $A,\ D_{ij},\ \partial D_{ij}/\partial x_i,\ V_i,\ \gamma,\ C^*,\ F\in H^{\alpha,\alpha/2}(\overline{Q}_T)$ for some $\alpha\in(0,1)$, $i,j=1,\cdots,n$.

$A(\cdot,0)$ is continuously differentiable in a neighborhood of S.

(2.21) $C_0\in W_q^{2-2/q}(\Omega)$ for some $q>(n+2)/2$, $q\geqq 2$, $q\neq n+2$.
C_0 is continuously differentiable in a neighborhood of S.
$N_0\in H^{\alpha}(\overline{\Omega})$.

(2.22) $C_*\in W_q^{1-1/q,(1-1/q)/2}(S_T)$.
$\beta,\ D_{ij}n_i\in H^{1-1/q+\varepsilon,(1-1/q+\varepsilon)/2}(\overline{S}_T)$ for some $\varepsilon>0$, $j=1,\cdots,n$.

(2.23) If $q>3$: $\partial C_0(x)/\partial\nu=\beta(x,0)(C_0(x)-C_*(x,0))$ for $x\in S$.

(2.24) $S\in H^{2+\alpha}$.

3. Existence of weak solutions. The differential equation in (2.14) can be written as

$$AC_t-LC+H(N)N_t=F, \tag{3.1}$$

where

$$H(s):=\begin{cases}1, & s>0,\\ 0, & s\leqq 0,\end{cases}\quad s\in\mathbb{R}. \tag{3.2}$$

Motivated by this, we develop the following notion:

DEFINITION 3.1. A pair $(C,\tilde{N})$ is called a *weak solution* in Q_T, if

$$C\in W_2^{1,1}(Q_T)\cap C(\overline{Q}_T),\ \tilde{N}\in C(\overline{Q}_T),\ C(x,0)=C_0(x) \text{ in } \overline{\Omega} \tag{3.3}$$

$$\begin{aligned}&\int_{Q_T}AC_t\phi\,dx\,dt+\int_{Q_T}D_{ij}\frac{\partial C}{\partial x_j}\frac{\partial\phi}{\partial x_i}dx\,dt\\ &\quad-\int_{S_T}\beta(C-C_*)\phi\,dx\,dt+\int_{Q_T}V_i\frac{\partial C}{\partial x_i}\phi\,dx\,dt\\ &\quad-\int_{Q_T}\hat{H}(\tilde{N})\phi_t\,dx\,dt-\int_{\Omega}\hat{H}(N_0(x))\phi(x,0)\,dx=\int_{Q_T}F\phi\,dx\,dt\end{aligned} \tag{3.4}$$

for all $\phi\in W_2^{1,1}(Q_T)$ such that $\phi(\cdot,T)=0$ a.e. in Ω, whereby $\hat{H}(s):=\max(s,0)$, $s\in\mathbb{R}$;

$$\tilde{N}(x,t)=N_0(x)-\int_0^t(\gamma(C^*-C))(x,\tau)\,d\tau \quad \text{in } \overline{Q}_T. \tag{3.5}$$

We first show the existence of a weak solution and later examine the relation between classical and weak solutions.

We use the following smoothing of H:

$$H_\varepsilon(s) := \begin{cases} 0, & s \leqq -\varepsilon/2, \\ s/\varepsilon + 1/2, & |s| < \varepsilon/2, \\ 1, & s \geqq \varepsilon/2 \end{cases} \tag{3.6}$$

and consider the following regularized problem:

DEFINITION 3.2. Let $\varepsilon > 0$. A pair $(C_\varepsilon, N_\varepsilon)$ is called a solution of (P_ε), if it classically fulfills

$$AC_{\varepsilon_t} - LC_\varepsilon + H_\varepsilon(N_\varepsilon) N_{\varepsilon_t} = F \quad \text{in } Q_T, \tag{3.7}$$

the identity (3.5), and the initial and boundary conditions (2.15).

LEMMA 3.3. (P_ε) *possesses a solution* $(C_\varepsilon, N_\varepsilon)$, $C_\varepsilon \in W_q^{2,1}(Q_T)$. *A constant* K *independent of* ε *exists such that*

$$\|C_\varepsilon\|^{(2)}_{q, Q_T} \leqq K. \tag{3.8}$$

Proof. Define the operator $\mathcal{F} = \mathcal{F}_\varepsilon$:

$$(\mathcal{F}(N))(x,t) := N_0(x) - \int_0^t (\gamma(C^* - C))(x,\tau)\, d\tau,$$

C being the classical solution $C_\varepsilon^{(N)}$ of

$$AC_t - LC - H_\varepsilon(N)\gamma(C^* - C) = F \quad \text{in } Q_T \tag{3.9}$$

with initial and boundary conditions (2.15).

Let $\delta := \min(\alpha, 2-(n+2)/q)/2$. We will prove the existence of a fixed point N_ε of $\mathcal{F}$ in $X := H^{\delta, \delta/2}(\overline{Q}_T)$. Obviously $(C_\varepsilon^{(N_\varepsilon)}, N_\varepsilon)$ is the desired solution. The proof will be accomplished by Schauder's theorem.

(i) $\mathcal{F}$ is well defined from X into X.

Let $N \in X$. $C_\varepsilon^{(N)}$ exists uniquely (e.g. [5, p. 147, Cor. 2]) and is in $W_q^{2,1}(Q_T)$ (compare [8, IV, §9], and [10, Thm. 17]); therefore $C_\varepsilon^{(N)}$ is also in $H^{\tilde{\alpha}, \tilde{\alpha}/2}(\overline{Q}_T)$, with $\tilde{\alpha} := 2-(n+2)/q$ [8, II, Lemma 3.3]. A fortiori we have $\mathcal{F}(N) \in X$.

In the following K denotes constants independent of N and ε:

ii) For some K the closed ball $B_K(0)$ is mapped by $\mathcal{F}$ into itself.

Let $N \in X$; then we have (compare [10, Thm. 17], [8, IV, §9])

$$\begin{aligned} \|C_\varepsilon^{(N)}\|^{(2)}_{q, Q_T} &\leqq K\left(1 + \|F + H_\varepsilon(N)\gamma(C^* - C_\varepsilon^{(N)})\|_{q, Q_T}\right) \\ &\leqq K\left(1 + \|C_\varepsilon^{(N)}\|_{q, Q_T}\right). \end{aligned}$$

We can conclude from the maximum principle (e.g. [8, Chap. I, Thm. 2.3])

$$|C_\varepsilon^{(N)}|^{(0)}_{Q_T} \leqq K,$$

and thus we get

$$\|C_\varepsilon^{(N)}\|^{(2)}_{q, Q_T} \leqq K \tag{3.10}$$

and especially ([8, Chap. II, Lemma 3.3])

$$|C_\varepsilon^{(N)}|_{Q_T}^{(2\delta)} \leqq K. \tag{3.11}$$

This also means that X is mapped into a ball $B_K(0)$.

iii) $\mathscr{F}$ is continuous.

Let $N_{1,2} \in X$; then $w := C_\varepsilon^{(N_1)} - C_\varepsilon^{(N_2)}$ fulfills

$$\begin{aligned} &Aw_t - Lw + H_\varepsilon(N_1)\gamma w = \gamma\big(C^* - C_\varepsilon^{(N_2)}\big)\big(H_\varepsilon(N_1) - H_\varepsilon(N_2)\big) := f \quad \text{in } Q_T, \\ &w(x,0) = 0, \qquad x \in \overline{\Omega}, \\ &\frac{\partial w}{\partial \nu} = \beta w \quad \text{on } S_T. \end{aligned}$$

Now applying [5, Thm. 4′, p. 213] we get:

$$|C_\varepsilon^{(N_1)} - C_\varepsilon^{(N_2)}|_{Q_T}^{(\delta)} \leqq K|f|_{Q_T}^{(0)} \leqq K/\varepsilon |N_1 - N_2|_{Q_T}^{(0)}$$

using e.g. (3.11). Therefore:

$$|\mathscr{F}(N_1) - \mathscr{F}(N_2)|_{Q_T}^{(\delta)} \leqq K/\varepsilon |N_1 - N_2|_{Q_T}^{(0)}.$$

To finish the proof, we have to establish:

iv) $\mathscr{F}[B_K(0)]$ is precompact, K according to ii).

This is true because of (3.11) and the compactness of the imbedding $H^{2\delta,\delta}(\overline{Q}_T) \to X$.

Finally due to (3.10) we also have (3.8). □

THEOREM 3.4. *There exists a weak solution* $(C, \tilde{N})$ *such that*

$$C \in W_q^{2,1}(Q_T).$$

Proof. The solutions of (P_ε) satisfy (3.5) and an integral relation $(3.4)_\varepsilon$, which equals (3.4) after substituting $\hat{H}$ by

$$\hat{H}_\varepsilon(s) := \int_0^s H_\varepsilon(\xi)\, d\xi. \tag{3.12}$$

Using (3.8), we see, by passing to a subsequence if necessary, that:

$$\begin{aligned} &C_\varepsilon \to C \quad \text{weakly in } W_q^{2,1}(Q_T) \qquad (3.13) \\ &C_\varepsilon \to C \quad \text{strongly in } H^{\tilde{\alpha},\tilde{\alpha}/2}(\overline{Q}_T) \qquad (3.14) \end{aligned} \quad \text{for } \varepsilon \to 0$$

with $\tilde{\alpha} \in (0, 2-(n+2)/q)$.

From (3.14) we conclude

$$N_\varepsilon \to \tilde{N} \quad \text{strongly in } C(\overline{Q}_T), \tag{3.15}$$

and $\tilde{N}$ fulfills (3.5). (3.3) is satisfied, too.

To finish the proof, we have to establish the convergence of $(3.4)_\varepsilon$ to (3.4), where due to (3.13) it suffices to show:

$$\gamma \int_{Q_T} \big(\hat{H}_\varepsilon(N_\varepsilon) - \hat{H}(\tilde{N})\big)\phi_t\, dx\, dt \to 0 \tag{3.16}$$

for $\varepsilon \to 0$ and all test functions ϕ.

We have:

$$(\hat{H}_\varepsilon(N_\varepsilon)-\hat{H}_\varepsilon(\tilde{N}))(x,t)=H_\varepsilon(\xi(x,t))(N_\varepsilon(x,t)-\tilde{N}(x,t))$$

for some $\xi(x,t)\in\langle N_\varepsilon(x,t),\ \tilde{N}(x,t)\rangle$, and therefore from (3.15)

$$|\hat{H}_\varepsilon(N_\varepsilon)-\hat{H}_\varepsilon(\tilde{N})|^{(0)}_{Q_T}\to 0 \quad \text{for } \varepsilon\to 0. \tag{3.17}$$

Furthermore $\|\hat{H}_\varepsilon-\hat{H}\|_{L^\infty}\to 0$ for $\varepsilon\to 0$ such that

$$|\hat{H}_\varepsilon(\tilde{N})-\hat{H}(\tilde{N})|^{(0)}_{Q_T}\to 0 \quad \text{for } \varepsilon\to 0. \tag{3.18}$$

(3.17), (3.18) imply (3.16). □

As a first regularity result let us mention:

THEOREM 3.5. *Let $(C,\tilde{N})$ be a weak solution such that $C\in W_p^{2,1}(Q_T)$ for some $p>1$.*

Let Q be a cylinder with smooth boundary, $\overline{Q}\subset Q_1$ or $\overline{Q}\subset Q_2$ with

$$Q_i:=\left\{(x,t)\in Q_T \mid (-1)^{i+1}\tilde{N}(x,t)>0\right\},\ i=1,2.$$

Then $C\in H^{2+\alpha,1+\alpha/2}(\overline{Q})$ and $(C,\tilde{N})$ fulfills classically

$$AC_t-LC+\tilde{N}_t=F \quad \text{in } Q_1, \tag{3.19}$$

$$QC_t-LC=F \quad \text{in } Q_2. \tag{3.20}$$

Proof. For definiteness we regard $\overline{Q}\subset Q_1$; the other case is strictly similar. Choose an analogous open cylinder $\tilde{Q}$ such that $\overline{Q}\subset\tilde{Q}\subset Q_1$. Let ϕ be an infinitely differentiable function in $\tilde{Q}$ with compact support. If we extend ϕ outside of $\tilde{Q}$ by 0, we obtain an admissible test function for (3.4) such that after partial integration we get

$$\int_{\tilde{Q}}(AC_t-LC+\tilde{N}_t-F)\phi\,dx\,dt=0$$

because $\hat{H}(\tilde{N})=\tilde{N}$ in $\tilde{Q}$. Therefore we have a generalized solution of (3.19) in $W_p^{2,1}(\tilde{Q})$, which by means of (3.5), possesses the desired regularity (compare [8, III, Thm. 12.2]). □

4. Uniqueness of weak and classical solutions. In this section we will use additionally the following:

Assumption 4.1.

1) $$A_t,\ \frac{\partial V_i}{\partial x_i}\in L^2(Q_T),\quad i=1,\cdots,n.$$

2) $$V_i n_i\in H^{1/2+\varepsilon,(1/2+\varepsilon)/2}(\overline{S}_T) \quad \text{for some } \varepsilon>0.$$

We start with the investigation of the following auxiliary problem for given f and h:

$$\left((A\phi)_t+\frac{\partial}{\partial x_i}\left(D_{ij}\frac{\partial\phi}{\partial x_j}\right)+\frac{\partial}{\partial x_i}(V_i\phi)\right)(x,t)$$
$$+\gamma(x,t)\int_t^T(h\phi_t)(x,\tau)\,d\tau=f(x,t) \quad \text{in } Q_T, \tag{4.1}$$

$$\phi(x,T)=0 \quad \text{in } \Omega, \tag{4.2}$$

$$\frac{\partial\phi}{\partial\nu}=(\beta-V_i n_i)\phi \quad \text{on } S_T. \tag{4.3}$$

LEMMA 4.2. *Let* $f \in L^2(Q_T)$, $h \in L^\infty(Q_T)$; *then there exists a generalized solution* ϕ *of* (4.1)–(4.3) *in* $W_2^{2,1}(Q_T)$.

Proof. Let $N \in \mathbb{N}$ and $\bar{T} := T/N$, $Q_k := Q_{T-k\bar{T},T-(k-1)\bar{T}}$ for $k = 1, \cdots, N$. For some $N \in \mathbb{N}$ we prove the existence of solutions $\phi_k \in W_2^{2,1}(Q_k)$ of

$$(4.4)\qquad \left((A\phi_k)_t + \frac{\partial}{\partial x_i}\left(D_{ij}\frac{\partial \phi_k}{\partial x_j}\right) + \frac{\partial}{\partial x_i}(V_i\phi_k)\right)(x,t) + \gamma(x,t)\int_t^{T-(k-1)\bar{T}} (h\phi_{k_t})(x,\tau)\,d\tau = f(x,t) - \gamma(x,t)\sum_{l=1}^{k-1}\int_{T-l\bar{T}}^{T-(l-1)\bar{T}} (h\phi_{l_t})(x,\tau)\,d\tau \quad \text{in } Q_k$$

with $\Sigma_{l=1}^{0} := 0$,

$$(4.5)\qquad \phi_k(x, T-(k-1)\bar{T}) = \phi_{k-1}(x, T-(k-1)\bar{T}) \quad \text{in } \Omega$$

with $\phi_{-1} := 0$, and of (4.3).

Let $\phi(x,t) := \phi_i(x,t)$ for $(x,t) \in Q_i$; we then get a function in $W_2^{2,1}(Q_T)$ because of (4.5), which is a solution of (4.1)–(4.3).

Now consider $1 \leqq k \leqq n$ and assume that the existence of a solution ϕ_l for $1 \leqq l < k$ is proven. For $k = 1$ this assumption is void. Define an operator $\mathcal{F}$ on $W_2^{2,1}(Q_k)$ by $\mathcal{F}z := u$, u being the solution in $W_2^{2,1}(Q_k)$ of (4.5), (4.3) and a differential equation similar to (4.4), but with

$$(4.6)\qquad \gamma(x,t)g^{(z)}(x,t), \quad g^{(z)}(x,t) := \int_t^{T-(k-1)\bar{T}} (hz_t)(x,\tau)\,d\tau,$$

as the integral term on the left-hand side.

In particular the right-hand side and $\gamma g^{(z)}$ are in $L^2(Q_i)$, and

$$\phi_{k-1}(\cdot, T-(k-1)\bar{T}) \in W_2^1(\Omega)$$

(e.g. [8, Chap. II, Lemma 3.4]); this operator is well defined (compare again [8, Chap. IV, §9], [10, Thm. 17]) and there exists a constant C independent of $\bar{T}$ and k such that for $z_1, z_2 \in W_2^{2,1}(Q_k)$:

$$(4.7)\qquad \|\mathcal{F}(z_1) - \mathcal{F}(z_2)\|_{2,Q_k}^{(2)} \leqq C\|g^{(z_1-z_2)}\|_{2,Q_k}.$$

An easy calculation shows for $z \in W_2^{2,1}(Q_k)$:

$$(4.8)\qquad \|g^{(z)}\|_{2,Q_k} \leqq |h|_{Q_T}^{(0)}\bar{T}\|z_t\|_{2,Q_k}.$$

Therefore it suffices to take N so large such that e.g.

$$(4.9)\qquad C|h|_{Q_T}^{(0)}\bar{T} \leqq \tfrac{1}{2}.$$

Then $\mathcal{F}$ is a contraction on $W_2^{2,1}(Q_k)$ and therefore Banach's fixed point theorem guarantees the unique existence of a solution ϕ_k in $W_2^{2,1}(Q_k)$ of (4.3)–(4.5). Thus we get a solution for all Q_k, which proves the assertion. □

Now we can show

LEMMA 4.3. *There exists at most a weak solution in* Q_T.

Proof. Let $(C,\tilde{N})$ be a weak solution and take test functions $\phi\in W_2^{2,1}(Q_T)$ such that $\phi(\cdot,T)=0$ a.e. in Ω. By partial integration we get:

$$
\begin{aligned}
(4.10)\quad & -\int_{Q_T} C(A\phi)_t\,dx\,dt-\int_\Omega (A\phi)(x,0)C_0(x)\,dx \\
& -\int_{Q_T} C\frac{\partial}{\partial x_i}\left(D_{ij}\frac{\partial\phi}{\partial x_j}\right)dx\,dt+\int_{S_T} CD_{ij}\frac{\partial\phi}{\partial x_j}n_i\,dx\,dt \\
& -\int_{S_T}\beta(C-C_*)\phi\,dx\,dt-\int_{Q_T} C\frac{\partial}{\partial x_i}(V_i\phi)\,dx\,dt \\
& +\int_{S_T} CV_in_i\phi\,dx\,dt-\int_{Q_T}\hat{H}(\tilde{N})\phi_t\,dx\,dt-\int_\Omega \hat{H}(N_0(x))\phi(x,0)\,dx \\
& =\int_{Q_T} F\phi\,dx\,dt.
\end{aligned}
$$

Now we consider two weak solutions $(C_i,\tilde{N}_i)$, $i=1,2$ and set $C:=C_1-C_2$, $\tilde{N}:=\tilde{N}_1-\tilde{N}_2$.

At first we show that

$$
(4.11)\qquad \hat{H}(\tilde{N}_1)-\hat{H}(\tilde{N}_2)=h(\tilde{N}_1-\tilde{N}_2)\quad\text{for some } h=h(\tilde{N}_1,\tilde{N}_2)\in L^\infty(Q_T).
$$

In fact, taking $Q_1:=\{(x,t)\in Q_T\mid(\tilde{N}_1-\tilde{N}_2)(x,t)\neq 0\}$ and $Q_2:=Q_T\backslash Q_1$, we can define h in the open set Q_1 as

$$
h(x,t):=\frac{(\hat{H}(\tilde{N}_1)-\hat{H}(\tilde{N}_2))(x,t)}{(\tilde{N}_1-\tilde{N}_2)(x,t)}.
$$

Thus h is a continuous function with values in $[0,1]$, which is easily checked. In Q_2 the definition is at our disposal, say $h(x,t):=0$. Therefore

$$
(4.12)\qquad (\hat{H}(\tilde{N}_1)-\hat{H}(\tilde{N}_2))(x,t)=h(x,t)\int_0^t(\gamma C)(x,\tau)\,d\tau
$$

and by partial integration:

$$
\begin{aligned}
(4.13)\quad & \int_{Q_T} h(x,t)\int_0^t(\gamma C)(x,\tau)\,d\tau\phi_t(x,t)\,dx\,dt \\
& =\int_{Q_T}(\gamma C)(x,t)\int_t^T(h\phi_t)(x,\tau)\,d\tau\,dx\,dt
\end{aligned}
$$

for all considered test functions ϕ. Let us apply (4.10) to C_1 and C_2 and subtract the two relations. By means of (4.12) and (4.13) we get:

$$
\begin{aligned}
(4.14)\quad & -\int_{Q_T} C(x,t)\left\{\left((A\phi)_t+\frac{\partial}{\partial x_i}\left(D_{ij}\frac{\partial\phi}{\partial x_j}\right)+\frac{\partial}{\partial x_i}(V_i\phi)\right)(x,t)\right. \\
& \qquad\left.+\gamma(x,t)\int_t^T(h\phi_t)(x,\tau)\,d\tau\right\}dx\,dt \\
& +\int_{S_T} C(x,t)\left(\frac{\partial\phi}{\partial\nu}-(\beta-V_in_i)\phi\right)(x,t)\,dx\,dt=0
\end{aligned}
$$

for all $\phi\in W_2^{2,1}(Q_T)$ such that $\phi(\cdot,T)=0$ a.e. in Ω.

Let $f \in L^2(Q_T)$ and take ϕ as a generalized solution of (4.1)–(4.3) with this f as right-hand side and $h = h(\tilde{N}_1, \tilde{N}_2)$. ϕ exists according to Lemma 4.2 and is an admissible test function. (4.14) implies

$$\int_{Q_T} C(x,t) f(x,t)\, dx\, dt = 0, \tag{4.15}$$

and thus $C = 0$, i.e. $C_1 = C_2$.

Because of (3.5) we also have $\tilde{N}_1 = \tilde{N}_2$, which proves our assertion. □

Remark 4.4. 1) We have actually proven that a solution $(C, \tilde{N})$, say $C, \tilde{N} \in C(\overline{Q}_T)$, of (4.10) and (3.5) is unique.

2) The additional assumption 4.1 should be as weak as possible so as not to be in conflict with the general situation of partially unsaturated soil.

THEOREM 4.5. *Let* (C_i, N_i, Φ_i) *be classical solutions,*

$$Q_+^i := \{ (x,t) \in Q_T \mid N_i(x,t) > 0 \}, \qquad Q_-^i := Q_T \backslash \overline{Q_+^i}, \qquad i = 1,2.$$

We assume:

$$\gamma(C^* - C_i) \geqq 0 \text{ on } Q_T, \ i = 1,2. \tag{4.16}$$

$$\text{The divergence theorem is valid in } Q_+^i \text{ and } Q_-^i, \qquad i = 1,2. \tag{4.17}$$

Then we have $(C_1, N_1, \Phi_1) = (C_2, N_2, \Phi_2)$.

Proof. Let (C, N, Φ) be a classical solution fulfilling (4.16), (4.17). Set

$$\tilde{N}(x,t) := N_0(x) - \int_0^t (\gamma(C^* - C))(x,\tau)\, d\tau. \tag{4.18}$$

Because of (4.16) we have

$$\tilde{N}(x,t) \leqq 0 \quad \text{for all } (x,t) \in Q_-. \tag{4.19}$$

In fact, (4.19) is implied by

$$\tilde{N}(x,t) > 0 \Rightarrow N(x,t) > 0 \quad \text{for } (x,t) \in Q_T. \tag{4.20}$$

To prove (4.20), suppose that $N(x,t) = 0$ for some $(x,t) \in Q_T$, for which $\tilde{N}(x,t) > 0$. Then $t \geqq \Phi(x)$ and therefore

$$0 < \tilde{N}(x,t) \leqq \tilde{N}(x, \Phi(x)). \tag{4.21}$$

On the other hand we can conclude

$$\tilde{N}(\xi,\tau) = N(\xi,\tau) \quad \text{for all } (\xi,\tau) \in \overline{Q}_+, \tag{4.22}$$

and thus

$$\tilde{N}(x, \Phi(x)) = N(x, \Phi(x)) = 0. \tag{4.23}$$

This is a contradiction to (4.21).

Now let $\phi \in C^{2,1}(\overline{Q}_T)$. We start with

$$\int_{Q_+} F\phi\, dx\, dt = \int_{Q_+} (AC_t - LC + N_t)\phi\, dx\, dt \tag{4.24}$$

and integrate by parts. Doing the same in Q_-, we can add these expressions and see that all integrals over the boundary $\Phi(x)=t$ cancel because of the continuity of N, C and $\partial C/\partial x_i$ in $\overline{Q}_T$. By means of (4.19) and (4.22) we can substitute for the integrals involving N

$$(4.25)\qquad -\int_{Q_T}\hat{H}(\tilde{N})\phi_t\,dx\,dt+\int_\Omega(\hat{H}(\tilde{N})\phi)(x,T)\,dx-\int_\Omega\hat{H}(N_0(x))\phi(x,0)\,dx.$$

The relation we now have achieved is also valid for all $\phi\in W_2^{2,1}(Q_T)$, as we see by passing to the limit. In this way we have verified (4.10) for $(C,\tilde{N})$. Therefore we see from Remark 4.4 that

$$(4.26)\qquad C_1=C_2,\qquad \tilde{N}_1=\tilde{N}_2,$$

$\tilde{N}_i$ being defined as in (4.18). It remains to prove $\Phi_1=\Phi_2$ and its consequence $N_1=N_2$.

Let $\Omega^i_{0,T}$ be the domains of Φ_i. At first we show:

$$(4.27)\qquad \Omega^1_{0,T}=\Omega^2_{0,T}=:\Omega_{0,T}.$$

Assume there exists $x\in\Omega^1_{0,T}$, such that $x\notin\Omega^2_{0,T}$. As

$$(4.28)\qquad \xi\notin\Omega_{0,T}\Rightarrow(\xi,\tau)\in\overline{Q}_+\quad\text{for all }\tau\in[0,T]$$

we have by (4.22)

$$(4.29)\qquad \tilde{N}_2(x,t)>0\quad\text{for all }t\in[0,T].$$

On the other hand by (4.23) $\tilde{N}_1(x,\Phi_1(x))=0$, which contradicts (4.26), (4.29).

Now let $x\in\Omega\cap\Omega_{0,T}$ and $\Phi_1(x)<\Phi_2(x)$. Then $N_2(x,\Phi_1(x))>0$, again leading to a contradiction by (4.22), (4.23). This proves $\Phi_1=\Phi_2$. $N_1=N_2$ follows along the same lines. □

Remark 4.6. As in the physical situation $\gamma>0$ and C^* is the saturation concentration, (4.16) is not restrictive (compare Theorem 5.5).

5. Existence of classical solutions. To obtain a classical solution from the weak solution constructed in Theorem 3.4, we must study the nature of the set $\tilde{N}=0$. This will be accomplished by deriving a lower bound for C^*-C. Since we cannot apply the strong maximum principle directly, we reconsider the solutions of the smoothed problem (P_ε). Throughout the section we will use

Assumption 5.1.

$$(5.1)\qquad \gamma(x,t)>0\quad\text{for all }(x,t)\in\overline{Q}_T.$$

$$(5.2)\qquad C^*\in H^{2+\alpha,1+\alpha/2}(\overline{Q}_T),$$
$$\hat{F}:=F-AC_t^*+LC^*\leqq 0\quad\text{in }\overline{Q}_T.$$

$$(5.3)\qquad \frac{\partial C^*}{\partial\nu}\geqq 0\quad\text{on }S_T.$$

$$(5.4)\qquad C_0(x)\leqq C^*(x,0)\quad\text{for all }x\in\overline{\Omega}.$$

(5.5) $$\beta \leqq 0 \qquad \text{on } S_T.$$

(5.6) $$C_* \leqq C^* \qquad \text{on } S_T.$$

(5.7) i) There exists $\bar{x} \in \Omega$ such that $C_0(\bar{x}) < C^*(\bar{x}, 0)$ or there exists a sequence $(x_n, t_n) \in S_T$, $t_n \to 0$, satisfying one of the following conditions:

ii) $\beta(x_n, t_n) < 0, \qquad C_*(x_n, t_n) < C^*(x_n, t_n),$

iii) $$\frac{\partial C^*}{\partial \nu}(x_n, t_n) > 0.$$

(5.8) $$N_0 \geqq 0 \quad \text{in } \bar{\Omega}.$$

Having in mind the discussion of the physical meaning in §2, we see that this leaching situation is covered by these assumptions. In particular, (5.7ii) is fulfilled, if at some portion of the boundary (the soil-atmosphere interphase) water, with a concentration beneath the saturation concentration, filters through.

As comparison function we regard the classical solution of

(5.9) $$Aw_t - Lw + \gamma w = -\hat{F} \quad \text{in } \bar{Q}_T,$$

(5.10) $$w(x,0) = C^*(x,0) - C_0(x), \qquad x \in \bar{\Omega},$$

(5.11) $$\frac{\partial w}{\partial \nu} = \beta\big(w - [C^* - C_*]\big) + \frac{\partial C^*}{\partial \nu} \quad \text{on } S_T.$$

w exists uniquely (e.g. [5, p. 147, Cor. 2]).

LEMMA 5.2. *$w(x,t) > 0$ for all $(x,t) \in \bar{\Omega} \times (0,T]$.*

Proof. Let us first show: $w \geqq 0$ in $\bar{Q}_T$. Suppose not; then w has a negative minimum in $(\bar{x}, \bar{t}) \in \bar{Q}_T$. Because of (5.1), (5.2) the strong minimum principle is applicable (e.g. [5, p. 34, Thm. 1]) and thus $(\bar{x}, \bar{t}) \in \Omega \times (0, T]$ contradicts (5.4).

The same is true for $\bar{t} = 0$. For $\bar{x} \in S$ we have, by a lemma of Viborny–Friedman ([5, p. 49, Thm. 14]), $(\partial w/\partial \nu)(\bar{x}, \bar{t}) < 0$ and thus using (5.3), (5.5), (5.6):

(5.12) $$w(\bar{x}, \bar{t}) > (C^* - C_*)(\bar{x}, \bar{t}) \geqq 0.$$

This is also a contradiction, i.e. $w \geqq 0$ in $\bar{Q}_T$.

If $w(\bar{x}, \bar{t}) = 0$ for some $(\bar{x}, \bar{t}) \in \bar{\Omega} \times (0, T]$, then w attains its minimum in $(\bar{x}, \bar{t})$. In the case $(\bar{x}, \bar{t}) \in \Omega \times (0, T]$ we get $w = 0$ in $\bar{Q}_{\bar{t}}$ by the strong minimum principle and thus a contradiction to (5.7). Finally, $\bar{x} \in S$ is impossible due to (5.12). □

LEMMA 5.3. *Let $(C_\varepsilon, N_\varepsilon)$ be a solution of* (P_ε). *Then*

$$C^* - C_\varepsilon \geqq w \quad \text{in } \bar{Q}_T.$$

Proof. Set $u_\varepsilon := C^* - C_\varepsilon$. Then u_ε solves

(5.13) $$Au_{\varepsilon_t} - Lu_\varepsilon + H_\varepsilon(N_\varepsilon)\gamma u_\varepsilon = -\hat{F} \quad \text{in } \bar{Q}_T$$

and (5.10), (5.11). Therefore $z := u_\varepsilon - w$ fulfills

(5.14) $$Az_t - Lz + \gamma H_\varepsilon(N_\varepsilon) z = \gamma\big(1 - H_\varepsilon(N_\varepsilon)\big) w \quad \text{in } \bar{Q}_T,$$

(5.15) $$z(x,0) = 0, \qquad x \in \bar{\Omega},$$

(5.16) $$\frac{\partial z}{\partial \nu} = \beta z \quad \text{in } S_T.$$

We want to show $z \geqq 0$ in $\bar{Q}_T$.

Since $\gamma H_\varepsilon(N_\varepsilon) \geqq 0$ and $\gamma(1-H_\varepsilon(N_\varepsilon))w \geqq 0$ due to Lemma 5.2, again we can apply the strong minimum principle to show this in the same way as in Lemma 5.2. □

We now conclude

THEOREM 5.4. *There exists a weak solution* $(C, \tilde{N})$, $C \in W_q^{2,1}(Q_T)$, *such that*

$$C(x,t) < C^*(x,t) \text{ for all } (x,t) \in \bar{\Omega} \times (0,T]. \tag{5.17}$$

$$\tilde{N}(x,\cdot) \text{ is strictly decreasing in } [0,T] \text{ for all } x \in \bar{\Omega}. \tag{5.18}$$

$$\begin{aligned} &\text{There is a } \psi \in C(\bar{\Omega}),\ 0 \leqq \psi(x) \leqq T \text{ for all } x \in \bar{\Omega} \text{ such that } \psi(x) > 0 \Leftrightarrow \\ &N_0(x) > 0 \text{ and for all } t \in [0, \psi(x)) : \tilde{N}(x,t) > 0, \text{ for all } t \in (\psi(x), T] : \\ &\tilde{N}(x,t) < 0,\ \psi(x) < T \Rightarrow \tilde{N}(x, \psi(x)) = 0. \end{aligned} \tag{5.19}$$

Proof. We consider the weak solution $(C, \tilde{N})$ constructed in Theorem 3.4. Let $\delta > 0$. Lemmas 5.2, 5.3 imply for some $K(\delta) > 0$

$$C^* - C_\varepsilon \geqq K(\delta) \quad \text{in } \bar{\Omega} \times [\delta, T],$$

and thus by (3.14) we have the validity of (5.17).

Furthermore, because of (5.1), we have for some $\underline{\gamma} > 0$:

$$N_{\varepsilon_i} \leqq -\underline{\gamma} K(\delta) \quad \text{in } \bar{\Omega} \times [\delta, T]$$

and therefore again by (3.14):

$$\tilde{N}_t \leqq -\underline{\gamma} K(\delta) \quad \text{in } \bar{\Omega} \times [\delta, T]. \tag{5.20}$$

In particular, this means (5.18). Now define

$$\psi(x) := \begin{cases} 0, & \tilde{N}(x,t) \leqq 0 \quad \text{for all } t \in [0,T], \\ \sup\{\, t \in [0,T] \mid \tilde{N}(x,t) > 0 \,\} & \text{otherwise.} \end{cases} \tag{5.21}$$

The properties in (5.19) are verified immediately. In particular, continuity is proved by checking upper and lower semicontinuity, which follows from the other assertions in (5.19). □

Collecting the results we get:

THEOREM 5.5. *Let* $q > n+2$. *There exists a classical solution* (C, N, Φ) *such that*

$$C \in W_q^{2,1}(Q_T) \text{ and } C \text{ fulfills } (5.17).$$

If $N_0 \in C^1(\bar{\Omega}_1)$ *and* γ *and* C^* *are continuously differentiable with respect to* x *in* $\bar{\Omega}_1 \times [0,T]$, *then*

$$\Phi \in C^1(\bar{\Omega}_1), \tag{5.22}$$

setting $\Omega_1 := \{ x \in \Omega_{0,T} \cap \Omega \mid 0 < \Phi(x) < T \}$.

Proof. We consider the weak solution $(C, \tilde{N})$ of Theorem 5.4. Defining $N(x,t) := (\tilde{N}(x,t))_+$ in $\bar{Q}_T$, we see

$$\begin{aligned} Q_+ &= \{\, (x,t) \in Q_T \mid \tilde{N}(x,t) > 0 \,\}, \\ Q_- &= \{\, (x,t) \in Q_T \mid \tilde{N}(x,t) < 0 \,\}. \end{aligned} \tag{5.23}$$

Therefore Theorem 3.5 assures (2.14), while (2.16), (2.17) are clear by the definition of N. Having in mind (3.14), we notice that (2.13) and (2.15) are satisfied.

We define $\Phi := \psi|_{\Omega_{0,T}}$, ψ according to Theorem 5.4. Then from (5.19) we conclude

$$\tilde{N}(x,\Phi(x))=0 \quad \text{for all } x\in\Omega_{0,T}, \tag{5.24}$$

and thus (2.11) and (2.12).

Finally (5.22) is a consequence of (5.24) and the implicit function theorem, the continuity of $\tilde{N}_x$ and $\tilde{N}_t$ in $\bar{\Omega}_1\times[0,T]$ and $\tilde{N}_t(x,\Phi(x))<0$ for $x\in\bar{\Omega}_1$. □

6. Asymptotic behavior. It is not yet clear whether N vanishes for large times, i.e. $\Phi(x)$ exists for each $x\in\bar{\Omega}$. We will give a sufficient condition for this situation. First, we regard all functions as given for all $T>0$ and the assumptions 2.2 and 5.1, which are also in force, are to be understood in this sense. Furthermore we use the following assumption in the sense of [5, p. 157] and with $Q := \bigcup_{T>0} Q_T$.

Assumption 6.1.

(6.1) $A(x,t)\geqq a$ for some $a>0$ and all $(x,t)\in\bar{Q}$,
$\gamma(x,t)\geqq \underline{\gamma}$ for some $\underline{\gamma}>0$ and all $(x,t)\in\bar{Q}$.

(6.2) $D_{ij}\xi_i\xi_j \geqq d|\xi|^2$ for some $d>0$ and all $(x,t)\in\bar{Q}$, $\xi\in\mathbb{R}^n$.

(6.3) The following functions converge as $t\to\infty$ uniformly in $\bar{Q}$ to a function in $H^\alpha(\bar{\Omega})$:

$$A,\ D_{ij},\ \sum_{i=1}^{n}\frac{\partial D_{ij}}{\partial x_i},\ V_i,\ \gamma,\ F,\ C_t^*,\ \frac{\partial C^*}{\partial x_i},\ \frac{\partial^2 C^*}{\partial x_i\partial x_j}, \qquad i,j=1,\cdots,n.$$

(6.4) $S=S_1\dot{\cup}S_2$, S_1 closed, and
$\beta(x,t)\leqq -b$ for some $b>0$ and all $x\in S_1$, $t>0$,
$\beta(x,t)\geqq -b$ for all $x\in S_2$, $t>0$.
$C^*(x,t)=C_*(x,t)$ for all $x\in\partial S_1$, $t>0$.

(6.5) $\beta(x,t)\to\beta(x)$, $C_*(x,t)\to C_*(x)$, $C^*(x,t)\to C^*(x)$ as $t\to\infty$ uniformly in S_1.

(6.6) There is an $\bar{x}\in S_1$ such that $C_*(\bar{x})<C^*(\bar{x})$.
In (5.7ii) the x_n are in S_1.

Again S_1 has to be understood as part of the soil-atmosphere interface and $-\beta$ as the flow velocity into the interior. With regard to the comparison function w it is necessary to change the boundary condition in its definition. To this end we extend β outside of S_1, such that the resulting function k has the properties

$$\begin{aligned} &k\in C(S_T),\ k_{|S_1\times[0,T]}=\beta \text{ for all } T>0,\\ &k(x,t)\leqq -b \text{ for all } (x,t)\in\bar{Q},\\ &k(x,t)\to k(x) \text{ as } t\to\infty \text{ uniformly in } S. \end{aligned} \tag{6.7}$$

Instead of (5.11) we now use as boundary condition for w

$$(6.8)\qquad \begin{aligned} &\frac{\partial w}{\partial \nu}=\beta\big(w-[C^*-C_*]\big)+\frac{\partial C^*}{\partial \nu} \quad \text{on } S_1\times[0,T],\\ &\frac{\partial w}{\partial \nu}=kw+\frac{\partial C^*}{\partial \nu} \quad \text{on } S_2\times[0,T]. \end{aligned}$$

The existence of a unique classical solution in Q_T for every $T>0$ is guaranteed. In the same manner as in §5 we prove the following lemmas.

LEMMA 6.2. *$w(x,t)>0$ for all $(x,t)\in\bar{\Omega}\times(0,T]$ for every $T>0$.*

LEMMA 6.3. *Let $(C_\varepsilon,N_\varepsilon)$ be a solution of (P_ε). Then $C^*-C_\varepsilon\geqq w$ in $\bar{Q}_T$ for every $T>0$.*

The conditions we have imposed are sufficient to insure the lower bound for w globally:

LEMMA 6.4. *For each $\delta>0$ there is a constant $K(\delta)>0$ such that $w(x,t)\geqq K(\delta)$ for all $x\in\bar{\Omega}$, $t\in[\delta,T]$ for every $T>\delta$.*

Proof. Assumption 6.1 assures that

$$(6.9)\qquad w(x,t)\to z(y),\qquad (x,t)\in\bar{Q} \quad \text{as } x\to y,\quad t\to\infty,$$

uniformly in $\bar{Q}$ ([5, p. 167, Thm. 5]), z being the solution of an elliptic problem with the limits of the coefficients and right-hand side as the corresponding terms. Applying the minimum principle and the lemma of Hopf (e.g. [5, p. 55, Thm. 21]) we can show using (6.6) and $\beta(\bar{x})<0$:

$$(6.10)\qquad z(x)>0 \quad \text{for all } x\in\bar{\Omega}.$$

From this, (6.9) and Lemma 6.2 the assertion follows. □

Before we can use this result together with Lemma 6.3, we have to assure that we can speak of $(C_\varepsilon,N_\varepsilon)$ in $\bar{Q}$ in a unique way:

LEMMA 6.5. *Let $T>0$. There exists at most a solution $(C_\varepsilon,N_\varepsilon)$ of (P_ε).*

Proof. Let $\varepsilon>0$ be fixed and (C_i,N_i), $i=1,2$, be solutions of (P_ε). $C:=C_1-C_2$ fulfills

$$(6.11)\qquad \begin{aligned} &(AC_t-LC+H_\varepsilon(N_1)\gamma C)(x,t)\\ &\qquad=\big(\gamma(C^*-C_2)B\big)(x,t)\int_0^t(\gamma C)(x,\tau)\,d\tau \quad \text{for } (x,t)\in Q_T, \end{aligned}$$

$$(6.12)\qquad C(x,0)=0,\qquad x\in\bar{\Omega},$$

$$(6.13)\qquad \frac{\partial C}{\partial \nu}=\beta C \quad \text{on } S_T,$$

whereby $B\in L^\infty(Q_T)$ such that in Q_T:

$$H_\varepsilon(N_1(x,t))-H_\varepsilon(N_2(x,t))=B(x,t)\int_0^t(\gamma C)(x,\tau)\,d\tau.$$

Now it is easy to show, with the aid of the usual L^∞-estimate (e.g. [8, Chap. I, Thm. 2.3]) applied to C and C_2 that for some $\bar{T}\in(0,T]$ the only possibility is $C\equiv 0$ in $\bar{Q}_{\bar{T}}$. In $Q_{\bar{T},2\bar{T}}$ we can repeat the argument and thus finally get $C\equiv 0$ in $\bar{Q}_T$, i.e. the uniqueness of the solution. □

THEOREM 6.6. *Let $q>n+2$. There exist (C,N,Φ) in $\bar{Q}$ and $\hat{T}>0$ such that (C,N,Φ) is a classical solution in Q_T for all $T>\hat{T}$ with the property $\Omega_{0,T}=\bar{\Omega}$, i.e. there is some $\hat{T}>0$ such that $N(x,t)=0$ for all $x\in\bar{\Omega}$, $t\geqq\hat{T}$.*

Proof. Because of Lemma 6.5 there are $(C_\varepsilon, N_\varepsilon)$ in $\overline{Q}$ such that their restrictions to Q_T are the solutions of (P_ε) in Q_T for every $T>0$. We can find a subsequence, denoted in the same way, such that

$$C(x,t)=\lim_{\varepsilon\to 0} C_\varepsilon(x,t), \qquad \tilde{N}(x,t)=\lim_{\varepsilon\to 0} N_\varepsilon(x,t) \quad \text{in } \overline{Q}$$

and in $\overline{Q}_T$ $(C,\tilde{N})$ is the weak solution constructed in Theorem 3.4.

As in the proof of Theorem 5.4, we can derive by using Lemmas 6.3, 6.4:

$$\tilde{N}_t \leqq -\underline{\gamma} K(\delta) \quad \text{in } \overline{\Omega}\times[\delta,T] \quad \text{for every } T>\delta. \tag{6.14}$$

Therefore there is a $\hat{T}>0$ such that

$$\tilde{N}(x,t)<0 \quad \text{for all } x\in\overline{\Omega}, \quad t\geqq\hat{T}. \tag{6.15}$$

Thus the solution constructed by Theorem 5.4 and 5.5 fulfills the assertion. □

Acknowledgments. This paper originated in part during the author's stay at the Istituto Matematico, University of Florence. The author would like to express his deep appreciation to Professors A. Fasano and M. Primicerio for their hospitality and their advice. Furthermore the author is indebted to Miss I. Kögel (University of Bayreuth) for introducing him to the field of soil science.

REFERENCES

[1] J. BEAR, *Hydraulics of Groundwater*, McGraw-Hill, New York, 1979.

[2] C. W. BOAST, *Modelling the movement of chemicals in soils by water*, Soil Sci., 151 (1973), pp. 224–230.

[3] E. BRESLER, B. L. MCNEAL, AND D. L. CARTER, *Saline and sodic soils, Principles-Dynamics-Modelling*, Springer, Berlin-Heidelberg-New York, 1982.

[4] FAO/UNESCO, *Irrigation, Drainage and Salinity—An International Source Book*, Hutchinson, London, 1973.

[5] A. FRIEDMAN, *Partial Differential Equations of Parabolic Type*, Prentice-Hall, Englewood Cliffs, NJ, 1964.

[6] V. E. KLYKOV, V. L. KULAGIN, AND V. A. MOROZOV, *The Stefan type problem occurring in the investigation of salt dissolution and transport process in soil*, Prikl. Matem. Mekh., 44 (1980), pp. 104–112. PMM USSR, 44 (1981), pp. 70–75.

[7] V. L. KULAGIN, *A parabolic problem with an unknown boundary*, Dokl. Akad. Nauk SSSR, 252 (1980), pp. 76–79. Sov. Phys. Dokl., 25 (1980), pp. 350–351.

[8] O. A. LADYZHENSKAJA, V. A. SOLONNIKOV, AND N. N. URAL'CEVA, *Linear and Quasilinear Equations of Parabolic Type*, American Mathematical Society, Providence, RI, 1968.

[9] D. MELAMED, R. J. HANKS, AND L. S. WILLARDSON, *Model of salt flow in soil with a source-sink term*, Soil Sci. Soc. Am. J., 41 (1971), pp. 29–33.

[10] V. A. SOLONNIKOV, *A priori estimates for second-order parabolic equations*, Trudy Mat. Inst. Steklov, 70 (1964), pp. 133–212; Amer. Math. Soc. Transl. Ser. 2, 65 (1967), pp. 51–138.

SIAM J. MATH. ANAL.
Vol. 17, No. 3, May 1986

008

A RELATION BETWEEN SEMI-INVERSE AND SAINT-VENANT SOLUTIONS FOR PRISMS*

DAVID KINDERLEHRER†

Abstract. It is shown that in an infinite prism the linearization at a natural state of a family of Ericksen's semi-inverse solutions is a combination of elementary St.-Venant solutions, namely, extension bending and torsion. Moreover, the span of these St. Venant solutions is precisely the linear manifold of solutions having locally uniformly bounded strain energy. This implies that any solution of a linearized problem in a finite prism continuable to a solution in the infinite prism in a manner that its energy on any portion of fixed length remains bounded, is an elementary St.-Venant solution.

AMS(MOS) subject classifications. Primary 35J65, 73C10, 73C50

Key words. nonlinear elliptic systems, asymptotic properties, finite elasticity, St.-Venant's principle

1. Introduction.

Consider an infinite prism

$$\mathbb{P}=\Omega\times\mathbb{R}\subset\mathbb{R}^3$$

where $\Omega\subset\mathbb{R}^2$ is a region with boundary Γ composed of a material with strain energy density $W(F)$, a real valued function of 3×3 matrices F with $\det F>0$. The body is in equilibrium in the configuration

$$y=y(x)\colon \mathbb{P}\to\mathbb{R}^3$$

provided

$$\delta\int_{\mathbb{P}} W(\nabla y)\,dx=0, \tag{1.1}$$

where this is taken in an appropriate sense. Here we mean that $y(x)$ is a smooth function satisfying the system of equations

$$\begin{aligned} &-\operatorname{div} S(\nabla y)=0 \quad \text{in } \mathbb{P},\\ &S(\nabla y)\nu=0 \quad \text{on } \partial\mathbb{P}=\Gamma\times\mathbb{R} \end{aligned} \tag{1.2}$$

where $S(F)=W'(F)=(\partial W/\partial F_j^i)$ is the Piola stress and ν is the outward directed normal of Γ.

Inasmuch as $\mathbb{P}$ is infinite and no boundary conditions prevail as $|x_3|\to+\infty$, one scarcely expects this problem to have a unique solution or even a finite dimensional manifold of them. The distinguished class discussed here are the semi-inverse solutions introduced by J. Ericksen [1], [2] and our intention is to illustrate their connections with the St.-Venant solutions familiar in the linear theory of elasticity. The St.-Venant solutions may be characterized in terms of energy, the subject of §2. They are also linearizations of semi-inverse solutions, which we show in §3. Existence of semi-inverse solutions and their properties will be considered elsewhere [5]. The St.-Venant flexure solutions arise only indirectly in the present context. They are disucssed in §4.

*Received by the editors July 30, 1984.
†School of Mathematics, University of Minnesota, Minneapolis, Minnesota 55455.

There are many ways of describing St. Venant solutions, some of which are in terms of energy (Sternberg and Knowles [11], Maisonneauve [9], Ericksen [3].) The one given here, although it may properly be interpreted as a technical device rather than a physical statement, is not a minimum principle but an extensibility property of the solution. This permits several ways of distinguishing St.-Venant and semi-inverse solutions; we may as well indulge ourselves in one of them whose origin is in the work of Ericksen [3], [4].

According to our theorem, the only solutions of the linearized problem of (1.2) in $\mathbb{P}$ with locally uniformly bounded strain energy are the St.-Venant solutions of extension, bending, and torsion. Incidentally they are also the ones with uniformly bounded strain, owing to the modern theory of elliptic systems. It follows that they are the only solutions in finite prisms which admit extensions to infinite ones retaining these properties. This renders them especially appropriate in a formulation of St.-Venant's principle and is one version of the set $S(\infty)$ defined in [4]. Moreover they are the linearizations of a family of finite deformations of the infinite prism $\mathbb{P}$.

Another interpretation is related more generally to nonlinear problems. In a suitable functional framework, the St.-Venant solutions are eigenfunctions with eigenvalue $\mu=0$ of a linearized problem. The existence of a family of solutions of the system (1.2) emanating from the eigenspace is strongly indicated under these circumstances. Semi-inverse solutions are on such family, a particular one, it is possible to show, whose (nonlinear) strain energy is uniformly bounded.

A different perspective is taken by Muncaster [8], [9] in his interesting work.

I take this opportunity to thank Professor Ericksen for introducing me to this question and for many stimulating discussions. I also thank Professor Muncaster for his generous remarks.

We introduce a few notions and state our results. Let $W(F)$ be a smooth function defined on 3×3 matrices $F=(F_j^i)$ with $\det F>0$ satisfying the condition of frame indifference

$$W(QF)=W(F) \quad \text{for } Q^TQ=\mathbb{1}, \quad \det Q=1,$$

and set

$$S(F)=W'(F)=\left(\partial W/\partial F_j^i\right)$$

its Piola–Kirchoff stress which satisfies, therefore,

$$S(QF)=QS(F) \quad \text{for } Q^TQ=\mathbb{1}, \quad \det Q=1. \tag{1.3}$$

We assume that the material is isotropic,[1]

$$S(\mathbb{1})=0,$$

and

$$S'[X]=\lim_{t\to 0}\frac{1}{t}S(\mathbb{1}+tX)=X+X^T+(a-1)\operatorname{tr}X\mathbb{1} \tag{1.4}$$

---

[1] St.-Venant solutions for linear isotropic materials are very familiar, thus our restriction here. Less well known is that they may be defined for any (homogeneous, let us say) material satisfying

$$S'[\varepsilon]\cdot\varepsilon\geqq\alpha_0|\varepsilon|^2, \quad \varepsilon=\varepsilon^T, \quad \alpha_0>0,$$

a fact known to St.-Venant. The conclusions of this work hold also in the anisotropic case. Additional discussion of this may be found in [5].

where $a > \frac{1}{3}$. Given a function $v = (v^1, v^2, v^3)$, we set

$$\varepsilon(v) = \tfrac{1}{2}(\nabla v + \nabla v^T) \quad \text{and} \quad \sigma(v) = S'[\nabla v] = S'[\varepsilon(v)].$$

The restriction on a ensures that there is an $\alpha > 0$ such that

$$\sigma(v) \cdot \varepsilon(v) \geqq \alpha |\varepsilon(v)|^2 \quad \text{for any } v \tag{1.5}$$

or simply that

$$S'[X] \cdot X = S'[X] \cdot \frac{1}{2}(X + X^T) \geqq \frac{\alpha}{4}(X + X^T)^2, \quad \text{for any matrix } X.$$

In this note, by a semi-inverse deformation we understand a smooth mapping

$$y: \overline{\mathbb{P}} \to \mathbb{R}^3$$

satisfying the condition of axial indifference

$$y(x', x_3 + t) = Q(t)(y(x) + p(t)), \qquad -\infty < t < \infty, \tag{1.6i}$$

for some rotation $Q(t), \det Q(t) = 1$, and vector valued function $p(t)$, which vary smoothly with t. We also impose the condition

$$\det \nabla y > 0 \quad \text{in } \mathbb{P} \tag{1.6ii}$$

Alternative formulations of (1.6) are suggested in the work of Ericksen and Muncaster. The first of these below is the one we shall use most frequently. The mapping $y(x)$ is a semi-inverse; deformation provided (1.6ii) is satisfied and there are a Λ, $\Lambda + \Lambda^T = 0$, a $\xi \in \mathbb{R}^3$, and a vector field

$$u: \overline{\Omega} \to \mathbb{R}^3$$

such that

$$\begin{aligned} &y(x) = R(x_3)(u(x') + p(x_3)), \qquad x \in \overline{\mathbb{P}}, \\ &R(x_3) = e^{x_3 \Lambda} \quad \text{and} \quad \frac{dp}{dx_3} + \Lambda p = \xi, \quad x_3 \in \mathbb{R}. \end{aligned} \tag{1.7}$$

or provided the Cauchy–Green strain

$$C = \nabla y^T \nabla y \quad \text{is a function of } x' \in \Omega. \tag{1.8}$$

Some clarification of this is given in [1], [5], but we point out that it is elementary. A semi-inverse solution is a semi-inverse deformation satisfying the field equations (1.2).

The four "elementary" St.-Venant solutions, corresponding to extension, bending with moments parallel to the x_1 and x_2 axes, and torsion, are denoted by v_0, v_1, v_2 and v_3 and are described in the appendix for the reader's convenience.

THEOREM 1. *Suppose that v, subject to the condition*

$$\int_{I_l} \int_{\Omega} |\varepsilon(v)|^2 dx \leqq M < \infty \tag{1.9}$$

where I_l is any interval of the x_3-axis of fixed length $2l > 0$, is a solution of the linear system

$$\begin{aligned} &-\operatorname{div} \sigma(v) = 0 \quad &&\text{in } \mathbb{P}, \\ &\sigma(v)\nu = 0 \quad &&\text{on } \partial\mathbb{P}. \end{aligned} \tag{1.10}$$

Then

$$v=\sum_0^3 \lambda_i v_i+\gamma$$

where γ *is an affine rigid motion and* $\lambda_i\in\mathbb{R}$, $i=0,1,2,3$. *Thus* v *is a linear combination of "elementary" St.-Venant solutions plus an affine rigid motion.*

Consider a family of semi-inverse solutions passing through $y=x$, which we abbreviate by writing "a family of semi-inverse solutions" and by which we intend a curve of semi-inverse solutions

$$\begin{aligned} &y=y(x,t), \qquad |t| \text{ small},\\ &y(x,0)=x,\\ &y(x,t)=R(x,t)(u(x',t)+p(x_3,t)),\\ &R(x_3,t)=e^{x_3 t\Lambda} \quad \text{for a fixed } \Lambda,\ \Lambda+\Lambda^T=0, \end{aligned} \tag{1.11}$$

and

$$\xi(t)=(1+t\lambda_0)e_3+t(\mu_1 e_1+\mu_2 e_2).$$

It should be clear that we have chosen R and ξ as the first order terms in t of an arbitrary smooth curve.

THEOREM 2. *Assume that* $y(x,t)$ *is a family of semi-inverse solutions, that is,* $y(x,t)$ *satisfies* (1.11) *and* (1.2). *Then*

$$\left.\frac{dy}{dt}\right|_{t=0}=\sum_0^3 \lambda_i v_i+\gamma \quad \text{in } \mathbb{P}$$

where the v_i *are the "elementary" St.-Venant solutions and* γ *is an affine rigid motion.*

The number of parameters available to a family of semi-inverse solutions are six, three from Λ and three from $\xi'(0)$. It is not difficult to check that different values of μ_1, μ_2 correspond to the slightly altered family

$$\tilde{y}(x,t)=R(x_3)Q(t)(u(x')+p(t))$$

for an appropriate family $Q(t)$,

$$Q(t)^T Q(t)=\mathbb{1}, \quad \det Q(t)=1, \quad Q(0)=\mathbb{1},$$

which leads to the conclusion that μ_1, μ_2 only determine a rigid motion and not a St.-Venant solution.

2. An energy characterization of St.-Venant solutions. This section is devoted to the proof of Theorem 1. We begin with an elementary lemma. For a displacement ζ, set

$$\sigma^3(\zeta)=\sigma(\zeta)e_3.$$

LEMMA 2.1. *Suppose that* $v=(v^1,v^2,v^3)$ *satisfies* (1.9) *and* (1.10). *Then for any affine rigid motion*

$$\int_{\Omega\times\{b\}} \sigma^3(v)\gamma\, dx' = \text{const.} \tag{2.1}$$

$$\int_{\Omega\times\{b\}} \sigma^3\left(\frac{\partial v}{\partial x_3}\right)\gamma\, dx' = 0, \tag{2.2}$$

for all $b\in\mathbb{R}$.

Proof. In view of (1.10) and the divergence theorem, for any interval $I=(a,b)$,

$$0=\int_{\Omega\times I}\operatorname{div}\sigma(v)\gamma\,dx=\int_{\Omega\times\{b\}}\sigma^3(v)\gamma\,dx'=\int_{\Omega\times\{a\}}\sigma^3(v)\gamma\,dx',$$

verifying (2.1). Writing

$$\int_{\Omega\times\{b\}}\sigma^3(v)\gamma\,dx'=\int_{\Omega}\sigma^3(v(x',b))\gamma(x',b)\,dx',$$

it follows that

$$0=\frac{\partial}{\partial b}\int_{\Omega\times\{b\}}\sigma^3(v)\gamma\,dx' \tag{2.3}$$

$$=\int_{\Omega\times\{b\}}\sigma^3\left(\frac{\partial v}{\partial x_3}\right)\gamma\,dx'+\int_{\Omega\times\{b\}}\sigma^3(v)\frac{\partial}{\partial x_3}\gamma\,dx'.$$

We want to prove that the second integral in (2.3) vanishes, which will require (1.9). With $\gamma(x)=c+\omega x$, $\omega+\omega^T=0$,

$$\frac{\partial}{\partial x_3}\gamma^i=\omega_{i3},\qquad \frac{\partial}{\partial x_3}\gamma=(\omega_{13},\omega_{23},0),$$

so

$$\int_{\Omega\times\{b\}}\sigma^3(v)\frac{\partial}{\partial x_3}\gamma\,dx'=\sum_{\mu<3}\int_{\Omega\times\{b\}}\sigma_{\mu3}(v)\omega_{\mu3}\,dx'$$

choose

$$\gamma^\mu=\omega_{\mu3}x_3,\qquad \gamma^3=\sum\omega_{3\mu}x_\mu=-\sum\omega_{\mu3}x_\mu.$$

Substituting this on (2.1),

$$b\int_{\Omega\times\{b\}}\sum_\mu\sigma_{\mu3}\omega_{\mu3}\,dx'-\int_{\Omega\times\{b\}}\sigma_{33}\sum\omega_{\mu3}x_\mu\,dx'$$

$$=a\int_{\Omega\times\{a\}}\sigma_{\mu3}\omega_{\mu3}\,dx'-\int_{\Omega\times\{a\}}\sigma_{33}\sum\omega_{\mu3}x_\mu\,dx'.$$

Again by (2.1),

$$\int_{\Omega\times\{a\}}\sigma_{\mu3}\omega_{\mu3}\,dx'=\int_{\Omega\times\{b\}}\sigma_{\mu3}\,dx'$$

so

$$(b-a)\int_{\Omega\times\{a\}}\sum\sigma_{\mu3}\omega_{\mu3}\,dx'=\int_{\Omega\times\{b\}}\sigma_{33}\sum\omega_{\mu3}x_\mu\,d\mu'-\int_{\Omega\times\{a\}}\sigma_{33}\sum\omega_{\mu3}x_\mu\,dx'.$$

Integrating this expression in x_3 over an interval of length l and then applying (1.9), we have

$$l(b-a)\int_{\Omega\times\{a\}}\sum\sigma_{\mu3}\omega_{\mu3}\,dx'=\int_{\Omega\times(b,b+l)}\sigma_{33}\sum\omega_{\mu3}x_\mu\,dx-\int_{\Omega\times(a,a+l)}\sigma_{33}\sum\omega_{\mu3}x_\mu\,dx$$

$$\leqq CM|\omega|.$$

Above we have also used the algebraic fact that

$$|\sigma(v)|^2 \leqq \text{const.}|\varepsilon(v)|^2.$$

Finally choosing

$$\omega_{\mu 3} = \operatorname{sgn} \int_{\Omega\times\{a\}} \sigma_{\mu 3}\, dx',$$

$$\sum \left| \int_{\Omega\times\{a\}} \sigma_{\mu 3}\, dx' \right| \leqq \frac{CM}{l(b-a)} \quad \text{for all } a<b.$$

Letting $b\to\infty$ gives that

$$\int_{\Omega\times\{a\}} \sigma_{\mu 3}\, dx' = 0, \qquad \mu = 1,2, \quad -\infty < a < \infty.$$

Recalling (2.3), the lemma is established. Q.E.D.

Before completing the proof of the theorem we wish to remark about the use of Korn's inequality. If u is a solution of

$$\begin{aligned} &-\operatorname{div}\sigma(u) = 0 \quad \text{in } \mathbb{P}, \\ &\sigma(u)\nu = 0 \quad \text{on } \partial\mathbb{P}. \end{aligned} \tag{2.4}$$

The well-known technique of difference quotients (Nirenberg [10]) yields that for any $l>0$ and interval $I_l(a)=\{|a-x_3|<l\}$,

$$\int_{\Omega\times I_l(a)} |\nabla^2 u|^2\, dx \leqq C \int_{\Omega\times I_{2l}(a)} \left(|\nabla u|^2 + |u|^2\right) dx, \qquad -\infty < a < \infty, \tag{2.5}$$

$$\leqq C \int_{\Omega\times I_{2l}(a)} \left(|\varepsilon(u)|^2 + |u|^2\right) dx$$

where C depends on l.

Let γ be an affine rigid motion chosen so that $\int_{\Omega\times I_{2l}(a)} \nabla(u-\gamma)\, dx$ is symmetric, so $u-\gamma$ also satisfies (2.4) and $\nabla^2 u = \nabla^2(u-\gamma)$. In addition, by Kron's inequality

$$\int_{\Omega\times I_{2l}(a)} |u-\gamma|^2\, dx \leqq C_k \int_{\Omega\times I_{2l}(a)} |\varepsilon(u-\gamma)|^2\, dx$$

$$= C_k \int_{\Omega\times I_{2l}(a)} |\varepsilon(u)|^2\, dx,$$

C_k = Korn's constant for $\Omega\times I_{2l}(a)$. Combining this with (2.5), we conclude that a solution u of the system (2.4) satisfies

$$\int_{\Omega\times I_l(a)} |\nabla^2 u|^2\, dx \leqq C_0 \int_{\Omega\times I_{2l}(a)} |\varepsilon(u)|^2\, dx \tag{2.6}$$

where C_0 depends on l but not a or u. In particular, if v satisfies (1.9), (1.10) then

$$\int_{\Omega\times I_l(a)} |\nabla^2 v|^2\,dx \leqq 2C_0M, \qquad -\infty < a < \infty. \tag{2.7}$$

LEMMA 2.2. *If v satisfies* (1.9) *and* (1.10), *then*

$$\frac{\partial}{\partial x_3}\varepsilon(v) = \varepsilon\left(\frac{\partial}{\partial x_3}v\right) = 0 \quad \text{in } \mathbb{P}.$$

It is well known that the only solutions of (1.10) whose strain is independent of the axial coordinate comprise the subspace spanned by $\{v_0, v_1, v_2, v_3\}$ and the affine rigid motions. Thus the lemma gives the theorem. For the reader's convenience, a brief description of this is given in the appendix.

Proof. Obviously

$$-\operatorname{div}\sigma(v_{x_3}) = 0 \quad \text{in } \mathbb{P},$$
$$\sigma(v_{x_3})\nu = 0 \quad \text{on } \partial\mathbb{P}.$$

One standard way to proceed is to multiply the above by ηv_{x_3} and integrate over $\mathbb{P}$ where, for $a < b$ and $l > 0$,

$$\eta = \eta(x_3) = \begin{cases} 1, & a \leqq x \leqq b, \\ 0, & x < a-2 \text{ or } x > b+l, \end{cases}$$

$\eta' = 1/l$ in $(a-l, a)$ and $\eta' = -1/l$ in $(b, b+l)$. Thus

$$\begin{aligned} 0 &= -\int_{\mathbb{P}} \operatorname{div}\sigma(v_{x_3})\eta v_{x_3}\,dx \\ &= \int_{\mathbb{P}} \sigma(v_{x_3})\cdot\nabla(\eta v_{x_3})\,dx \\ &= \int_{\mathbb{P}} \eta\sigma(v_{x_3})\cdot\varepsilon(v_{x_3})\,dx + \int_{\mathbb{P}} \sigma(v_{x_3})\cdot\nabla\eta\otimes v_{x_3}\,dx \\ &= \int_{\mathbb{P}} \eta\sigma(v_{x_3})\cdot\varepsilon(v_{x_3})\,dx + \int_{\mathbb{P}} \sigma(v_{x_3})\cdot\eta' e_3\otimes v_{x_3}\,dx. \end{aligned}$$

Since $\sigma\cdot\varepsilon \geqq \alpha|\varepsilon|^2$,

$$\begin{aligned} \alpha\int_{\Omega\times(a,b)} |\varepsilon(v_{x_3})|^2\,dx &\leqq -\int_{\mathbb{P}} \sigma(v_{x_3})\cdot\eta' e_3\otimes v_{x_3}\,dx \\ &= \frac{1}{l}\int_{\Omega\times(b,b+l)} \sigma(v_{x_3})\cdot v_{x_3}\,dx - \frac{1}{l}\int_{\Omega\times(a-l,a)} \sigma(v_{x_3})\cdot v_{x_3}\,dx. \end{aligned}$$

Thus for some $C > 0$,

$$\int_{\Omega\times(a,b)} |\varepsilon(v_{x_3})|^2\,dx \leqq C\left\{\int_{\Omega\times(b,b+l)} \sigma^3(v_{x_3})v_{x_3}\,dx - \int_{\Omega\times(a-l,a)} \sigma^3(v_{x_3})v_{x_3}\,dx\right\}.$$

According to Lemma 2.1, for any affine rigid motions γ, β the right-hand side is unaltered if v_{x_3} is replaced by $v_{x_3}-\gamma$ or $v_{x_3}-\beta$. Consequently, using (2.7),

$$\begin{aligned}\int_{\Omega\times(a,b)} |\varepsilon(v_{x_3})|^2\,dx &\leqq C\Big\{\int_{\Omega\times(b,b+l)} \sigma^3(v_{x_3})(v_{x_3}-\gamma)\,dx \\ &\qquad -\int_{\Omega\times(a-l,a)} \sigma^3(v_{x_3})(v_{x_3}-\beta)\,dx\Big\} \\ &\leqq C\Big\{\|\nabla^2 v\|_{L^2(\Omega\times I_l(b))}\|v_{x_3}-\gamma\|_{L^2(\Omega\times I_l(b))} \\ &\qquad +\|\nabla^2 v\|_{L^2(\Omega\times I_l(a))}\|v_{x_3}-\beta\|_{L^2(\Omega\times I_l(a))}\Big\} \\ &\leqq CC_0M^{1/2}\Big(\|v_{x_3}-\gamma\|_{L^2(\Omega\times I_d l(b))}+\|v_{x_3}-\beta\|_{L^2(\Omega\times I_l(a))}\Big).\end{aligned}$$

Choosing γ and β appropriately, by Korn's inequality we conclude that

$$\begin{aligned}(2.8)\quad \int_{\Omega\times(a,b)} |\varepsilon(v_{x_3})|^2\,dx &\leqq 2CC_0M^{1/2}\Big\{\|\varepsilon(v_{x_3})\|_{L^2(\Omega\times I_l(b))}+\|\varepsilon(v_{x_3})\|_{L^2(\Omega\times I_l(a))}\Big\} \\ &\leqq 4CC_0M\end{aligned}$$

so that $\varepsilon(v_{x_3})\in L^2(\mathbb{P})$. Again, with l fixed, choose sequences $a_k\to-\infty$ and $b_k\to+\infty$ such that

$$\|\varepsilon(v_{x_3})\|_{L^2(\Omega\times I_l(b_k))}+\|\varepsilon(v_{x_3})\|_{L^2(\Omega\times I_l(a_k))}\to 0.$$

From (2.8) we conclude that

$$\int_{\mathbb{P}} |\varepsilon(v_{x_3})|^2\,dx=0$$

so $\varepsilon(v_{x_3})=0$ in $\mathbb{P}$. Q.E.D.

3. Semi-inverse solutions. A brief presentation of the field equations for a semi-inverse solution is our starting point. Assume that

$$(3.1)\qquad y(x)=R(x_3)(u(x')+p(x_3)),\qquad x\in\mathbb{P}$$

is a semi-inverse deformation. Temporarily writing $e=e_3$,

$$\begin{aligned}\nabla y(x) &= R(x_3)\Big(\nabla u+\frac{dp}{dx_3}\otimes e\Big)+R(x_3)\Lambda(u+p)\otimes e \\ &= R(x_3)(\nabla u+\Lambda u\otimes e+\xi\otimes e) \\ &\equiv R(x_3)U(x')\end{aligned}$$

and

$$S(\nabla y)=RS(U).$$

It follows that if y is a semi-inverse solution, then u is a solution of

$$(3.2)\qquad \begin{aligned}&\operatorname{div} S(U)+\Lambda S^3(U)=0 \quad \text{in } \Omega,\\ &S(U)\nu=0 \quad \text{on } \Gamma,\\ &U=\nabla u+\Lambda u\otimes e+\xi\otimes e,\end{aligned}$$

where $S^3(F)=S^3(F)e_3$ denotes the third column of $S(F)$. No ambiguity in (3.2) concerning the symbol "div" is possible since U depends only on x'.

Restricting our attention to deformations $y(x,t)$ satisfying (1.11), which entails imposition of initial conditions at $t=0$, we are able to write a full set of field equations for u and p. So we have for $|t|$ small

$$\begin{aligned} &U=U'(x',t)=\nabla u+t\Lambda u\otimes e+\xi(t),\\ &U(x',0)=\mathbb{1} \end{aligned} \tag{3.3}$$

satisfies

$$\begin{aligned} &\operatorname{div} S(U)+t\Lambda S^3(U)=0 \quad &&\text{in } \Omega,\\ &S(U)\nu=0 &&\text{on } \Gamma \end{aligned} \tag{3.4}$$

and

$$\begin{aligned} &\frac{dp}{dx_3}+t\Lambda p=\xi(t), \qquad -\infty<x_3<\infty,\\ &p(x_3,0)=x_3e,\\ &p(0,t)=0. \end{aligned} \tag{3.5}$$

Observe that

$$\left.\frac{dy}{dt}\right|_{t=0}=\left(\frac{du}{dt}+\frac{dp}{dt}\right)\Big|_{t=0}+x_3\Lambda x. \tag{3.6}$$

An incomplete version of Theorem 2 may be easily derived. Since

$$W(\nabla y)=W(RU)=W(U)$$

and U is a function of x', for small $|t|$ there is a constant $C(t)$, varying smoothly with t, such that

$$\int_{\Omega\times I_l(a)} W(\nabla y)\,dx=C(t)l, \qquad -\infty<a<\infty.$$

Thus

$$\int_{\Omega\times I_l(a)} S(\nabla y)\left[\nabla\frac{dy}{dt}\right]dx=\frac{d}{dt}\int_{\Omega\times I_l(a)} W(\nabla y)\,dx=c'(t)l,$$

also independent of $a\in\mathbb{R}$. By Taylor's theorem, with

$$\begin{aligned} v&=\left.\frac{dy}{dt}\right|_{t=0},\\ S(\nabla y)&=S(\mathbb{1})+t\nabla v+O(t^2))\\ &=S(\mathbb{1})+tS'(\mathbb{1})[\nabla v]+O(t^2)\\ &=t\sigma(v)+O(t^2) \end{aligned}$$

since $S(\mathbb{1})=0$. Consequently v is a solution of (1.10) satisfying

$$\int_{\Omega\times I_l(a)} \sigma(v)\cdot\varepsilon(v)\,dx=c_1l \qquad -\infty<a<\infty,$$

which implies (1.9). By Theorem 1 we conclude that

$$\left.\frac{dy}{dt}\right|_{t=0} = v = \sum_0^3 \mu_i v_i + \gamma \quad \text{in } \mathbb{P},$$

which in some qualitative way is the content of Theorem 2. However, it does not identify the coefficients μ_i with the matrix Λ and λ_0, so we turn instead to the explicit linearization of the equations (3.4), (3.5). This will also illustrate that the complete four-parameter family of St.-Venant solutions may be realized as linearizations of semi-inverse ones, provided, of course, the latter exist.

The linearization will give a system of equations for

$$\zeta = \left.\frac{du}{dt}\right|_{t=0}.$$

Indeed, since $U = \mathbb{1}$ at $t=0$,

$$\frac{dU}{dt} = \nabla\zeta + \Lambda u_0 \otimes e + \xi'(0) \otimes e,$$

$$u_0(x') = \begin{pmatrix} x_1 \\ x_2 \\ 0 \end{pmatrix},$$

we obtain

$$\operatorname{div} S'\left[\frac{dU}{dt}\right] + \Lambda S^3(\mathbb{1}) = 0 \quad \text{in } \Omega,$$

$$S'\left[\frac{dU}{dt}\right]\nu = 0 \quad \text{on } \Gamma$$

where $S' = S'(\mathbb{1})$. Now $S(\mathbb{1}) = 0$ gives

$$\begin{aligned} -\operatorname{div}\sigma(\zeta) &= \operatorname{div} S'[\Lambda u_0 \otimes e + \xi'(0) \otimes e] = f \quad \text{in } \Omega, \\ \sigma(\zeta)\nu &= -S'[\Lambda u_0 \otimes e + \xi'(0) \otimes e] = g \quad \text{on } \Gamma. \end{aligned} \tag{3.7}$$

With $'$ denoting the first two components, it may be convenient to note that (3.7) may be rewritten

$$\begin{aligned} -\Delta\zeta' - a\nabla \operatorname{div}\zeta' &= f' \quad \text{in } \Omega, \\ \sigma(\zeta')\nu &= g' \quad \text{on } \Gamma \end{aligned} \tag{3.8$_1$}$$

and

$$\begin{aligned} -\Delta\zeta^3 &= f_3 \quad \text{in } \Omega, \\ \frac{\partial\zeta^3}{\partial\nu} &= g_3 \quad \text{on } \Gamma. \end{aligned} \tag{3.8$_2$}$$

In addition

$$q(x_3) = \left.\frac{dp}{dt}\right|_{t=0}$$

is the solution of

$$\frac{dq}{dx_3} = \xi'(0) - x_3\Lambda e, \qquad q(0) = 0. \tag{3.9}$$

The systems (3.7), (3.9) are linear in λ_0 and the elements of Λ so it suffices to consider four separate cases corresponding to one $\lambda_i = 1$ with the rest set to zero. This may seem presumptuous, but we reserve to the end discussion of the bogus parameters μ_1, μ_2 which occur in $\xi'(0)$.

Set

$$\Lambda = \begin{pmatrix} 0 & -\lambda_3 & \lambda_2 \\ \lambda_3 & 0 & -\lambda_1 \\ -\lambda_2 & \lambda_1 & 0 \end{pmatrix}.$$

and suppose $\mu_1 = \mu_2 = 0$.

Case 0. (extension) $\lambda_0 = 1$, $\lambda_i = 0$, $i = 1,2,3$. Here the system (3.7) is

$$\begin{aligned} &-\operatorname{div}\sigma(\zeta) = 0 \quad \text{in } \Omega, \\ &\sigma(\zeta)\nu = -S'[e \otimes e]\nu = (1-a)\nu \quad \text{on } \Gamma \end{aligned}$$

and, clearly

$$q(x_3) = x_3 e.$$

So $\zeta = (1-a)/2a u_0(x') = -c_p u_0(x')$, where

$$c_p = \frac{1}{2a}(a-1) \tag{3.10}$$

denotes the Poisson ratio. Thus

$$\left.\frac{dy_0}{dt}\right|_{t=0} = v_0 = (-c_p x_1, -c_p x_2, x_3).$$

We turn now to torsion

Case 3. (torsion) $\lambda_3 = 1$, $\lambda_i = 0$, $i = 0,1,2$. In terms of (3.8), the system is

$$\begin{aligned} &-\Delta\zeta' - a\nabla \operatorname{div}\zeta' = 0 && \text{in } \Omega, \\ &\sigma(\zeta')\nu = 0 && \text{on } \Gamma, \\ &-\Delta\zeta^3 = 0 && \text{in } \Omega, \\ &\frac{\partial\zeta^3}{\partial\nu} = x_2\nu_1 - x_1\nu_2 && \text{on } \Gamma. \end{aligned}$$

Hence $\zeta' = 0$ and $\zeta^3 = \varphi$, the St.-Venant warping function. Moreover, $q \equiv 0$. Hence

$$\left.\frac{dy_3}{dt}\right|_{t=0} = v_3 = (-x_2x_3, x_1x_3, \varphi(x')), \qquad x \in \mathbb{P}.$$

Case 1. (e_1-bending) $\lambda_1 = 1$, $\lambda_i = 0$, $i = 0,2,3$. In this case

$$\Lambda u_0 = x_2 e$$

for which it is easily checked that

$$\begin{aligned} &-\Delta\zeta' - a\nabla \operatorname{div}\zeta' = \operatorname{div}(a-1)x_2\mathbb{1} = \begin{pmatrix} 0 \\ a-1 \end{pmatrix} && \text{in } \Omega, \\ &\sigma(\zeta')\nu = -(a-1)x_2\nu && \text{on } \Gamma \end{aligned}$$

and

$$\Delta\zeta^3=0 \quad \text{in } \Omega,$$
$$\frac{\partial\zeta^3}{\partial\nu}=0 \quad \text{on } \Gamma.$$

A solution is given by

$$\zeta(x')=\left(-c_p x_1 x_2, -\frac{1}{2}c_p\left(x_2^2-x_1^2\right),0\right).$$

For the auxiliary function q we find that

$$q(x_3)=+\frac{1}{2}x_3^2 e_2.$$

Combining these in (3.6) gives that

$$\frac{dy_1}{dt}\bigg|_{t=0}=v_1=\left(-c_p x_1 x_2, -\frac{c_p}{2}\left(x_2^2-x_1^2\right)-\frac{1}{2}x_3^2, x_2 x_3\right) \quad \text{on } \mathbb{P} \tag{3.11}$$

with c_p given by (3.10).

Finally, assume that all $\lambda_i=0$, $i=1,2,3$, and μ_i are given. For ζ this leads to the equations

$$-\operatorname{div}\sigma(\zeta)=0,$$
$$\sigma(\zeta)\nu=-\mu\cdot\nu e$$

so $\zeta=(0,0,-\mu\cdot x')$ while

$$\frac{dq}{dx_3}=\mu_i e_i, \qquad q(0)=0$$

implies $q(x_3)=x_3\mu_i e_i$. Thus

$$\frac{dy}{dt}\bigg|_{t=0}=\gamma x=\begin{pmatrix}0 & 0 & \mu_1\\ 0 & 0 & \mu_2\\ -\mu_1 & -\mu_2 & 0\end{pmatrix}x,$$

a rigid motion.

4. St.-Venant's flexure solution. The St.-Venant flexure solutions, which do not satisfy the energy restriction (1.9), are not easily accomodated in the present framework. In this way our discussion is more restrictive than Muncaster's [8], [9]. The flexure solution, however, is not especially satisfactory, even in the linear theory [2], [3]. We show here that they may be interpreted, rather casually, as displacements close to the linearized solutions of semi-inverse ones. Additional comments follow.

First consider the semi-inverse e_1-bending solution, corresponding to case 1 of §3, whose form is

$$\eta(x)=R(x_3)\left(u(x')+p(x_3)\right) \quad \text{in } \mathbb{P}$$

for

$$R(x_3)=e^{\delta x_3 M_1}, \quad M_1=\begin{pmatrix}0 & 0 & 0\\ 0 & 0 & -1\\ 0 & 1 & 0\end{pmatrix}, \quad \delta \text{ small},$$

and suppose there exists a family of solutions of the field equations (1.2) of the type

$$y(x,t)=Q(t)\eta(x,t), \qquad |t| \text{ small}$$

with $\eta(x,0)=\eta(x)$ as above and

$$Q(t)=e^{tx_3H}, \qquad H=\begin{pmatrix}0 & -1 & 0\\ 1 & 0 & 0\\ 0 & 0 & 0\end{pmatrix}.$$

Calculation of $dy/dt|_{t=0}$ and approximation of the linearized equations at $\eta(x,0)$ by those at $y=x$, (1.10), leads to the credible approximation of $dy/dt|_{t=0}$ by the St.-Venant flexure solution.

Another interpretation is in terms of second derivatives. Suppose now that

$$y=y(x,\mu,\lambda)=R(x_3)(u(x')+p(x_3)) \quad \text{in } \mathbb{P}$$

is a two parameter family of semi-inverse solutions with

$$\Lambda=\begin{pmatrix}0 & -\lambda & 0\\ \lambda & 0 & -\mu\\ 0 & \mu & 0\end{pmatrix}.$$

Assuming that

$$\frac{\partial^2 y}{\partial\mu\,\partial\lambda}\bigg|_{\mu=\lambda=0}$$

may be written as flexure + error, the error contributes no resultant force on the planes $x_3=\text{const}$. Actually this is wrong by a numerical coefficient. Moreover, the term "flexure" here means just that portion of the St.-Venant solution involving the classical flexure functions.

The suggested conclusion, which in view of the brevity of this discussion, the reader may be reluctant to embrace, is that St.-Venant flexure is sufficiently close to the semi-inverse family that it does not determine, by itself, an especially interesting family of solutions of the finite equations. Geometrically, analogues of large twisting and bending may be achieved by seeking solutions

$$y(x)=e^{x_3\Lambda}(u(x')-b+p(x_3))$$

subject to the conditions $p(0)=0$ and

$$\int_\Omega u(x')\,dx=0,$$

as discussed in Ericksen [2].

5. Some remarks. From Theorem 1, the kernel of a certain linear operator defined on the unbounded domain $\mathbb{P}$ is identified. For example, the Banach space may be chosen to be

$V=$ functions which are locally in $C^{2,\alpha}(\overline{\mathbb{P}})$ with $\|\varepsilon(v)\|_{C^{0,\alpha}(\overline{\mathbb{P}})}+\|\nabla^2 v\|_{C^{0,\alpha}(\overline{\mathbb{P}})}<\infty$,

$Y=C^{0,\alpha}(\mathbb{P})\times C^{1,\alpha}(\partial\mathbb{P})$

with the operator

$$L: V \to Y, \qquad Lv = (-\operatorname{div}\sigma(v), \sigma(v)\nu).$$

It would be useful to know the index of L, especially for the study of small deformations superimposed on large. This is also connected to the notion of stability of solutions.

Given a family of semi-inverse solutions of the form

$$y = y(x) = y(x, \Lambda, \lambda_0) = R(x_3)(u(x') + p(x_3)), \qquad x \in \mathbb{P},$$
$$R(x_3) = e^{x_3 \Lambda}$$

depending on four parameters $\lambda_0, \cdots, \lambda_3$, let us say, the displacements $\partial y/\partial \lambda_i$, $i = 0, 1, 2, 3$, are good candidates for St.-Venant solutions associated to the solution $y(x)$. From the standpoint of energy, these linearizations do not seem to have any apparent interpretation, such as that given by Theorem 1. The argument at the beginning of §3 relies on the vanishing of the stress tensor at $\mathbb{1}$. The linearized strain, however, E, given by

$$2E = \left(\nabla \frac{\partial y}{\partial \lambda_i}\right)^T F + F^T \left(\nabla \frac{\partial y}{\partial \lambda_i}\right), \qquad F = \nabla y,$$
$$= \frac{\partial}{\partial \lambda_i} U^T U$$

is independent of x_3. Is every solution of the linearized equations at $y = y(x)$ with this property such a "St.-Venant" solution?

Appendix. St.-Venant solutions. For the reader's convenience we take up here the description of displacements $v(x)$ with

$$\frac{\partial}{\partial x_3} \varepsilon(v) = 0 \quad \text{in } \mathbb{P} \tag{A.1}$$

which implies that

$$\frac{\partial v}{\partial x_3} = \Lambda x + c, \quad \Lambda + \Lambda^T = 0, \quad c \in \mathbb{R}^3. \tag{A.2}$$

As the expression above is linear in the parameters occurring in Λ and c, it suffices to consider separate cases and to then superimpose their solutions. We consider here the case of torsion, where

$$\Lambda = H = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \qquad c = 0.$$

Solutions $u = (u^1, u^2)$ to the two-dimensional system

$$-\operatorname{div}\sigma(u) = 0 \quad \text{in } \Omega,$$
$$\sigma(u)\nu = 0 \quad \text{on } \Gamma$$

are unique to within a affine rigid motion. Our choice of St.-Venant solution will reflect this property.

From (A.2),

$$\frac{\partial v^1}{\partial x_3} = -x_2, \quad \frac{\partial v^2}{\partial x_3} = x_1, \quad \frac{\partial v^3}{\partial x_3} = 0$$

whence

$$\begin{aligned} v^1 &= -x_2 x_3 + \varphi^1(x^1), \\ v^2 &= x_1 x_3 + \varphi^2(x') \quad \text{in } \mathbb{P}, \\ v^3 &= \varphi^3(x'). \end{aligned}$$

Substituting this in the equations (1.10) gives

$$\begin{aligned} &\operatorname{div} \sigma(v) = \Delta\varphi + a\nabla \operatorname{div}\varphi = 0 \quad \text{in } \mathbb{P}, \\ &\sigma(v)\nu = 0 \quad \text{on } \partial\mathbb{P} \end{aligned}$$

so, with $\varphi' = (\varphi^1, \varphi^2)$,

$$\begin{aligned} &\Delta\varphi' + a\nabla \operatorname{div}\varphi' = 0 \quad \text{in } \Omega, \\ &\sigma(\varphi')\nu = 0 \quad \text{on } \Gamma \end{aligned}$$

and

$$\begin{aligned} &\Delta\varphi^3 = 0 \quad \text{in } \Omega, \\ &\frac{\partial \varphi^3}{\partial \nu} + (-x_2\nu_1 + x_1\nu_2) = 0 \quad \text{on } \Gamma. \end{aligned}$$

Thus φ' is a two-dimensional rigid motion we may set equal to zero and

$$\varphi^3 = \varphi$$

the St.-Venant warping function. Thus

$$v = (-x_2 x_3, x_1 x_3, \varphi(x')), \qquad x \in \mathbb{P}.$$

The bending and extension solutions may be determined in the same way. We do not pursue this here since they are described, from a different viewpont, in §3.

Of course, Love [6] is an ideal source of information about St.-Venant solutions.

REFERENCES

[1] J. L. ERICKSEN, *Special topics in elastostatics*, Adv. Appl. Mech., 17 (1977), pp. 189–243.

[2] _____, *On the formulation of St.-Venant's problem*, in Nonlinear Analysis and Mechanics: Heriot–Watt Symposium, Vol. I, R. J. Knops, ed., Pitman, London, 1977, pp. 158–186.

[3] _____, *On the status of St.-Venant's solutions as minimizers of energy*, Int. J. Solids Structures, 16 (1980), pp. 195–198.

[4] _____, *St.-Venant's principle for elastic prisms*, in Systems of Nonlinear PDE, J. M. Ball, ed., Oxford Univ. Press, Cambridge, 1983, pp. 87–93.

[5] D. KINDERLEHRER, *Some remarks about the existence of semi-inverse solutions*, to appear.

[6] A. E. H. LOVE, *A Treatise on the Mathematical Theory of Elasticity*, Dover, New York, 1948.

[7] O. MAISONNEUVE, There, Poitiers, 1971.

[8] R. MUNCASTER, *St. Venant's problem in nonlinear elasticity: a study of cross-sections*, in Nonlinear Analysis and Mechanics: Heriot–Watt Symposium, Vol. IV, R. J. Knops, ed., Pitman, London, 1979, pp. 17–75.

[9] _____, *St.-Venant's problem for slender prisms*, Utilitas Mathematica, to appear.

[10] L. NIRENBERG, *On elliptic partial differential equations*, Ann. S. N. S. Pisa, 13 (1959), pp. 1–48.

[11] E. STERNBERG AND J. KNOWLES, *Minimum energy characterization of St. Venant's solution to the relaxed St.-Venant's problem*, Arch., Rat. Mech. Anal., 21 (1966), pp. 89–107.

SIAM J. MATH. ANAL.
Vol. 17, No. 3, May 1986

009

ON THE BEHAVIOR OF THE SOLUTIONS TO THE LAMM EQUATION OF THE ULTRACENTRIFUGE II*

ATSUSHI YOSHIKAWA†

Abstract. In the ultracentrifugal analysis of a two-component solute-solvent system, the concentration $c(r,t)$ of the solute is described by the Lamm equation $\partial c/\partial t = r^{-1}\partial(r\{D_0\partial c/\partial r - r\omega^2 s_0 c/(1+kc)\})/\partial r$, $0<r_a<r<r_b$, $t>0$, with the nonlinear boundary conditions $D_0\partial c/\partial r - r\omega^2 s_0 c/(1+kc)=0$ at $r=r_a$ and $r=r_b$, and the initial data $c=c_0(r)$ when $t=0$. Here D_0, s_0, k, ω are positive constants independent of c. In the present paper, it is shown that for any smooth nonnegative initial data $c_0(r)$, compatible with the boundary conditions, the solution $c(r,t)$ converges to the equilibrium solution as $t\to+\infty$. The rate of convergence is also given. These improve some of the results obtained in [SIAM J. Math. Anal., 15 (1984), pp. 686–711] under more stringent assumptions on $c_0(r)$.

Introduction. Consider the following Lamm equation of the ultracentrifugal analysis:

$$\frac{\partial c}{\partial t} = r^{-1}\frac{\partial}{\partial r}\left\{r\left(D_0\frac{\partial c}{\partial r} - r\omega^2 sc\right)\right\}, \tag{0.1}$$

$0<r_a<r<r_b$, $t>0$, with the (nonlinear) boundary conditions

$$D_0\frac{\partial c}{\partial r} - r\omega^2 sc = 0 \tag{0.2}$$

at $r=r_a$ and $r=r_b$, and the initial data

$$c = c_0(r) \quad \text{when } t=0. \tag{0.3}$$

Here $c(r,t)$ stands for the concentration of the solute in a two-component system (see Fujita [1]). D_0 is the diffusion coefficient and ω the frequency of the rotor. We assume that D_0 and ω are positive constants independent of c. s is the sedimentation coefficient, which is given by

$$s = s(c) = \frac{s_0}{1+kc} \tag{0.4}$$

with positive constants s_0 and k.

From the physical point of view, it is natural to require that the initial concentration $c_0(r)$ satisfy

$$c_0(r) \geqq 0 \tag{0.5}$$

and

$$\int_{r_a}^{r_b} c_0(r)\,r\,dr > 0. \tag{0.6}$$

* Received by the editors May 12, 1983, and in revised form September 15, 1984.

† Department of Mathematics, Faculty of Science, Hokkaido University, Sapporo 060, Japan. Present address, Department of Applied Science, Faculty of Engineering, Kyushu University, Hakozaki, Fukuoka 812, Japan.

Some of the basic properties of the solution $c(r,t)$ of the Lamm equation (0.1)–(0.6) have been established in [2]. In particular, the existence and uniqueness of the solution to the Lamm equation have been verified [2, Thm. 1]. The property of the equilibrium solution $c_E(r)$ has also been discussed [2, Thm. 2]. Namely, $c_E(r)$ satisfies the equation

$$D_0\frac{\partial c_E}{\partial r}-r\omega^2 s(c_E)c_E=0 \tag{0.7}$$

in $r_a<r<r_b$ with

$$\int_{r_a}^{r_b} c_E(r)r\,dr=m_0. \tag{0.8}$$

However, the convergence of $c(r,t)$ to $c_E(r)$ as $t\to\infty$ has been shown only under the requirement that either $c_0(r)$ be nondecreasing in r or $c_0(r)$ satisfy

$$\left|D_0\frac{\partial c_0(r)}{\partial r}-r\omega^2 s(c_0(r))c_0(r)\right|\leqq\frac{D_0^2}{s_0\omega^2 r_b^3 k}$$

[2, Thm. 3].

In the present paper, we remove these restrictions on $c_0(r)$ and prove the following:

THEOREM 0.1. *Let $c_0(r)$ be a smooth function, compatible with the boundary conditions* (0.2), *and satisfying* (0.5) *and* (0.6). *Let $c(r,t)$ be the corresponding classical solution to the problem, and $c_E(r)$ be the equilibrium solution with the same mass as $c_0(r)$. Then there is a positive constant C independent of t, r such that for each $t>0$,*

$$|c(\cdot,t)-c_E|_\infty\leqq C\exp(-\lambda_E t). \tag{0.9}$$

In the above statement, $|\ |_\infty$ stands for the norm of the Banach space $L^\infty(r_a,r_b)$. λ_E is the smallest positive eigenvalue of the linearization $\hat{L}_E$ at $c=c_E$. That is, $\hat{L}_E$ is the self-adjoint extension in the Hilbert space $L^2(r_a,r_b;q(r)r\,dr)$ of the operator L_E:

$$L_E u=-r^{-1}\frac{\partial}{\partial r}\left(r\left\{D_0\frac{\partial}{\partial r}u-rs_0\omega^2 s_1'(c_E(r))u\right\}\right)$$

for $u\in C^2(r_a,r_b)$ with

$$D_0\frac{\partial}{\partial r}u-rs_0\omega^2 s_1'(c_E(r))u=0$$

at $r=r_a$ and $r=r_b$. Here we have employed $s_1(c)=c/(1+kc)$ and $s_1'(c)=1/(1+kc)^2$.

$$q(r)=\exp\left\{-D_0^{-1}s_0\omega^2\int_{r_a}^{r}s_1'(c_E(r'))r'\,dr'\right\}.$$

For the benefit of the readers, we include a few words about how $L_E u$ is derived. Suppose a solution c of (0.1)–(0.2) is written in the form $c=c_E+u$. Then from (0.1) and (0.2), we get

$$\frac{\partial u}{\partial t}-r^{-1}\frac{\partial}{\partial r}\left(r\left\{D_0\frac{\partial}{\partial r}u-rs_0\omega^2 s_1'(c_E(r))u\right\}\right)=\frac{1}{r}\frac{\partial}{\partial r}R(r;u,c_E),$$

$t>0$, $r_1<r<r_b$, and

$$D_0\frac{\partial}{\partial r}u-rs_0\omega^2 s_1'(c_E(r))u=R(r;u,c_E)$$

at $r=r_a$ and $r=r_b$. Here

$$R(r;u,c_E)=-\frac{r\omega^2 s_0 k u^2}{(1+kc_E)(1+kc_E+ku)}$$

and is considered to be of second order in u because of the fact that $c_E>0$, $c_E+u\geqq 0$ [2, Thm. 1]. Thus to obtain the part of order 1 in u, we put $R=0$ and get the operator L_E.

The proof of Theorem 0.1 is given in the next section. It depends on a better choice of the Lyapunov functional. Our technique relies on smoothness of the solution $c(r,t)$. In the meantime, Professor Barbu has shown me that the above problem (0.1)–(0.4) falls in the category of problems to which the monotone operator theory (in $L^2(r_a,r_b;r\,dr)$) is applicable. The initial concentration may then merely be in $L^2(r_a,r_b;r\,dr)$. Since such generalized possibly multivalued solutions are out of the scope of the techniques employed in the present note, we do not know if Theorem 0.1 still prevails under such a weak requirement on the initial concentration. However, it is certainly the case once the solution achieves regularity at some instant.

1. Proof of Theorem 0.1. Let $c(r,t)$ be the smooth classical solution of the problem (0.1)–(0.4) with the initial data satisfying (0.5) and (0.6). The smoothness and compatibility requirements of the initial concentration $c_0(r)$, on the other hand, are to guarantee the existence of such a smooth solution.

Let

$$J(c(r,t))=D_0\frac{\partial c}{\partial r}-rs_0\omega^2 s_1(c) \tag{1.1}$$

for $c=c(r,t)$. Recall that $s_1(c)=c/(1+kc)$. To control the behavior of $c(r,t)$ as $t\to+\infty$, we employ the following Lyapunov functional:

$$E(c(\cdot,t))=\int_{r_a}^{r_b} J(c(r,t))^N r\,dr, \tag{1.2}$$

with N a positive even integer large enough.

We can then show:

PROPOSITION 1.1. *Given nonnegative smooth compatible initial data $c_0(r)$, there are an even positive integer N_0 and a positive number q such that for any even integer $N\geqq N_0$, we have*

$$E(c(\cdot,t))\leqq e^{-qt}E(c_0), \qquad t>0.$$

Once this proposition is admitted, it is fairly immediate that the ω-limit set $\omega_p(c_0)$ of $c_0(r)$ in $L^p(r_a,r_b;r\,dr)$, $1\leqq p<\infty$, or in $C^0(r_a,r_b)$ (i.e., $p=\infty$), consists of the unique point $c_E(r)$, the equilibrium solution with the mass $m_0=\int_{r_a}^{r_b}c_0(r)r\,dr$. In fact, there is a sequence $t_1<t_2<\cdots<t_n<\cdots\to+\infty$ such that $J(c(r,t_i))$ converges to 0 almost everywhere in (r_a,r_b). Then the proof of [2, Thm. 3] is applicable to derive (0.9).

Proof of Proposition 1.1. Because of our regularity assumptions, $u(r,t)=J(c(r,t))$ satisfies the following equation:

$$\frac{\partial u}{\partial t}=D_0\frac{\partial^2 u}{\partial r^2}+A\frac{\partial}{\partial r}u-Bu, \tag{1.3}$$

$r_a<r<r_b$, $t>0$, with the boundary condition

$$u=0 \quad \text{at } r=r_a \text{ and } r=r_b, \tag{1.4}$$

and the initial data

$$u=J(c_0(r)) \quad \text{when } t=0. \tag{1.5}$$

Here

$$A=A(r,c)=\frac{D_0}{r}-s_0\omega^2 s_1'(c)r,$$

$$B=B(r,c)=s_0\omega^2 s_1'(c)+\frac{D_0}{r^2},$$

with $s_1'(c)=1/(1+kc(r,t))^2$ (see [2, §3]). Since

$$B\geqq\frac{D_0}{r_a^2}>0, \tag{1.6}$$

we have

$$|u(r,t)|\leqq\sup_r J(c_0(r))=M_0$$

in view of the maximum principle. Since $s_1(c)\leqq 1/k$, we have

$$\left|\frac{\partial c(r,t)}{\partial r}\right|\leqq D_0^{-1}M_0+r_b s_0\omega^{k-1}=M_1 \tag{1.7}$$

for all $t>0$ and $r_a<r<r_b$. Now using (1.3) and (1.4), we get by a routine computation

$$\frac{d}{dt}E(c(\cdot,t))=-N(N-1)D_0\int_{r_a}^{r_b}u^{N-2}u_r^2 r\,dr-\int_{r_a}^{r_b}(rA)_r u^N dr-N\int_{r_a}^{r_b}Bu^N r\,dr. \tag{1.8}$$

By (1.7),

$$(rA_r)=A+rA_r=\frac{D_0}{r}-3s_0\omega^2 r s_1'(c)-s_0\omega^2 r^2 s_1''(c)c_r \tag{1.9}$$

$$\geqq\frac{D_0}{r_b}-3s_0\omega^2 r_b-2ks_0\omega^2 r_b^2 M_1=M_2.$$

Therefore, from (1.6), (1.8) and (1.9) we get

$$\frac{d}{dt}E(c(\cdot,t))\leqq -qE(c(\cdot,t)) \tag{1.10}$$

if $N\geqq N_0>-M_2 r_a^2/(D_0 r_b)$, and thus

$$q=N\frac{D_0}{r_a^2}+\frac{M_2}{r_b}>0.$$

Proposition 1.1 is now immediate from (1.10) and (1.5).

Observe that the exponent q in Proposition 1.1 depends on N and also on $c_0(r)$. On the other hand, we know that

$$\lim_{p\to\infty}\left\{\int_{r_a}^{r_b}|v(r)|^p r\,dr\right\}^{1/p}=|v|_\infty$$

for any bounded measurable function $v=v(r)$ on the interval (r_a, r_b). From this fact, we immediately get the following:

COROLLARY 1.2. *Under the assumptions of Proposition* 1.1 *we have*

$$|J(c(\cdot,t))|_\infty \leqq e^{-Qt}|J(c_0(\cdot))|_\infty, \qquad t>0,$$

with $Q \geqq D_0/r_a^2$.

It is also immediate that $|J(c(\cdot,t))|_\infty$ is a decreasing function of t. This shows that $|J(c(\cdot,t))|_\infty$ would be the right choice as the Lyapunov functional, although the functional $E(c(\cdot,t))$ of (1.2) and Proposition 1.1 already provide enough information for our present purpose.

REFERENCES

[1] H. FUJITA, *Foundations of Ultracentrifugal Analysis*, John Wiley, New York, 1975.

[2] A. YOSHIKAWA, *On the behavior of the solutions of the Lamm equation of the ultracentrifuge*, this Journal, 15 (1984), pp. 686–711.

SIAM J. MATH. ANAL.
Vol. 17, No. 3, May 1986

010

R-SEPARATION FOR HEAT AND SCHRÖDINGER EQUATIONS I *

GREGORY JOHN REID†

Abstract. We classify all R-separable coordinate systems for the equation

$$(*) \qquad \Delta_m \Psi + 2\varepsilon \partial_t \Psi = E\Psi, \qquad \Delta_m = \sum_{u=1}^{m} \partial_{y^u y^u},$$

which for $\varepsilon = \frac{i}{2}$ and $\varepsilon = -\frac{1}{2}$, $E=0$ yields standardised versions of the Schrödinger and Heat equations respectively. The classification problem is solved by utilising the fact that (*) is a symmetry-reduced version of the Helmholtz equation on $m+2$ dimensional Minkowski space for which there is a well developed theory of separation. In all cases the separated solutions are eigenfunctions of a commuting set of operators which are at most quadratic in the operators which generate the motions of the symmetry group of (*). A detailed treatment is provided of the physically interesting case $m=3$.

1. Introduction. In this article a complete treatment of the variable separation problem is given for the equation

$$(*) \qquad \Delta_m \Psi + 2\varepsilon \partial_t \Psi = E\Psi \qquad \left(\Delta_m = \sum_{u=1}^{m} \partial_{y^u y^u}\right),$$

in the real variables y^u and t. When $\varepsilon = i/2$ this equation is a standardised version of the "constant potential Schrödinger equation" which will be referred to as the "Schrödinger equation". If $\varepsilon = -\frac{1}{2}$ and $E=0$ then (*) is a standardised version of the Heat or diffusion equation.

The coordinate system

$$(1.1) \qquad \begin{aligned} &y^u = y^u(\mathbf{x}), \qquad u = 1,2,\cdots,m, \\ &t = t(\mathbf{x}), \qquad \mathbf{x} = (x^1,\cdots,x^{m+1}), \end{aligned}$$

is *R-separable* if there are complex analytic functions Ψ, Ψ_j and R, such that (*) admits a solution of form

$$(1.2) \qquad \Psi(\mathbf{x},\mathbf{c}) = e^{R(\mathbf{x})} \prod_{j=1}^{m+1} \Psi_j(x^j,\mathbf{c}),$$

where $\mathbf{c} = (c_1,\cdots,c_{m+1})$ are the $m+1$ separation constants. It is the main task of this article to classify these systems. At first it might seem that any solution of (*) is R-separable but the independence of R from the constants c_j severely limits the possible R-separable solutions. Pure separation corresponds to $R=0$ and trivial R-separation to

$$(1.3) \qquad R = \sum_{i=1}^{m+1} R_i(x^i).$$

In the time independent case (i.e. $\partial_t \Psi = 0$), equation (*) reduces to the Helmholtz equation if $E \neq 0$ and to the Laplace equation when $E=0$ on real Euclidean space $\mathbb{R}^m$.

* Received by the editors January 30, 1984, and in revised form May 15, 1984.

† Mathematics Department, University of Waikato, Hamilton, New Zealand. Present address, Mathematics Department, South Dakota State University, Brookings, South Dakota 57007-1297.

A classification of the R-separable systems for these equations in three spatial dimensions ($m=3$) was first provided by Böcher (1894). Eisenhart (1934) gave a rigorous derivation of the systems obtained by Böcher. Eisenhart (1934) also characterized variable separation in a geometric manner that could be systematically applied to the variable separation problem on any Riemannian space.

Separable solutions $\Phi=\Pi_1^m\Phi_i(x^i)$ of the time independent form of (*) $\Delta_m\Phi=K\Phi$, also provide separable solutions $\Psi=e^{\lambda t}\Phi$ of the time-dependent form of (*) if $K+2\varepsilon\lambda=E$. However there are many R-separable systems for (*) where the time dependence is not so trivial. Only recently has a systematic investigation of such systems been undertaken. Kalnins and Miller (1974) classified all R-separable systems for the potential-free Schrödinger equation in one spatial dimension (i.e. (*) with $\varepsilon=i/2$ and $m=1$). Subsequently Boyer, Kalnins and Miller (1975a) classified all R-separable systems for the Schrödinger equation in two spatial dimensions. These and earlier investigators showed that the separability of these equations was intimately related to their respective point symmetry groups. More specifically they showed that the R-separated solutions were eigenfunctions of a set of mutually commuting sets of linear and/or quadratic combinations of the partial differential operators that generate the motions of the group. In Boyer, Kalnins and Miller (1975a, b) these group theoretical characterisations are used to give efficient and well-motivated derivations of addition theorems both old and new for the various special functions that appeared as eigensolutions of the separated problems. In his book *Symmetry and Separation of Variables*, Miller (1977) collects the results of the above papers and also characterises group theoretically variable separation for many of the common partial differential equations of mathematical physics.

The present paper is an extension of Kalnins and Miller's work to m dimensions. Kalnins and Miller (1982a) have already constructed all separable systems for $\Delta_m\Psi=E\Psi$ (i.e. (*)with $\varepsilon=0$). It is out of their construction that we will be able to build all R-separable systems for the time dependent case.

The *symmetry group* (Bluman and Cole (1974), Boyer (1974)) of (*) will prove to be of great use in our study. This is the group of coordinate transformations which leaves the form of (*) invariant. If two R-separable systems are related by the action of this group then we will not distinguish between them. We will often work with the *Lie* algebra corresponding to this group. The infinitesimal generators

$$L=\alpha^i(\mathbf{Y})\partial_{Y^i}+b(\mathbf{Y}) \tag{1.4}$$

for this algebra are found by solving the relation

$$[Q,L]=M(\mathbf{Y})Q, \tag{1.5}$$

where $Q=\Delta_m+2\varepsilon\partial_t-E$ and $\mathbf{Y}=(y^1,\cdots,y^m,t)$. Here [,] is the commutator bracket and we have used the Einstein summation convention which will always be assumed unless indicated otherwise. These generators can also be interpreted as symmetry operators in the sense that they map solutions to solutions: if Φ is a solution of (*) then so too is $L\Phi$.

Another illustration of the importance of the symmetry group is in the case of *ignorable variables*. A variable is *ignorable* if it is possible to choose (via the freedom (1.8a) defined below) an R-separable coordinate system such that this variable does not appear explicitly in (*). Greek indices will be used to denote such variables. Nonignorable coordinates will be referred to as being *essential variables*. Ignorables correspond to elements of the Lie algebra and systems containing N of them correspond to N

dimensional abelian subalgebras. The determination of these subalgebras under the adjoint action of the Lie algebra will be of great use in the solution of our problem.

In all cases the R-separated solutions Ψ will be characterised as eigenfunctions of a commuting set of $m+1$ partial differential operators quadratic in the elements of the Lie algebra. We will thus be able to extend the group theoretical characterisation found in lower dimensions by Kalnins and Miller (1974), Shapovalov and Sukhomlin (1974) to the time dependent Heat and Schrödinger equations of any spatial dimension.

The partial differential operators corresponding to ignorables are first order so the eigenfunction condition implies that

$$L_\alpha \Psi = l_\alpha \Psi, \tag{1.6}$$

in which case the x^α dependence in Ψ is simply

$$\Psi_\alpha = e^{l_\alpha x^\alpha}. \tag{1.7}$$

The remainder of the operators are second order.

To simplify this study we will say that two R-separable coordinate systems $\{\mathbf{x}\}$ and $\{\bar{\mathbf{x}}\}$ are *equivalent* if

$$\bar{x}^a = f^a(x^a), \tag{1.8a}$$

$$\bar{x}^\alpha = c^\alpha_\beta x^\beta + \sum_a g_a(x^a), \qquad \det(c^\alpha_\beta) \neq 0. \tag{1.8b}$$

or

(1.8c) they are related by the action of the symmetry group.

Here the c^α_β are real constants and the x^a are essential variables. Benenti (1980) has given a rigorous definition of *equivalence* for the Hamilton–Jacobi equation. By his definition two separable coordinate systems for this equation are equivalent if their separated complete integrals are the same (see §2 for the definition of separation for the Hamilton–Jacobi equation). He shows that by using this definition the only possible types of equivalence transformations are (1.8a) and (1.8b). These two equivalences extend naturally to our case since both of these transformations preserve the R-separability of (∗). We add (1.8c) because equivalence under the group removes the distinction between many different coordinate transformations leading to the same functional form of (∗).

In the early stages of this work a program (see Reid (1984)) was written in the symbolic language MACSYMA capable of producing all the time-consuming details of separation for flat spaces. Although a general theory was finally set up dispensing with the need for this program, it proved extremely useful in checking the results.

The structure of this article is as follows. In §2 we transform (∗) to the Helmholtz equation in Minkowski space, and exploit the fact that both this equation and the even simpler Hamilton–Jacobi equation separate in the same coordinate systems. Section 2 is also introductory in nature as it outlines much of the basic material needed for our study. In §3 we determine the possible sets of commuting Killing vectors (or abelian sub algebras) characterising separable systems for the Hamilton–Jacobi equation. Using these abelian subalgebras a simple form is then found for the metric. As a result the Riemann curvature conditions are solved in §4 and the metric completely determined. The technical work of finding the coordinate transformations is also carried out in §4. This work is summarised in graphical form in §5. In §6 we determine the operators for (∗) and in §7 the R-separable solutions. The results for physically interesting cases $m = 1, 2$, and 3 are tabulated in Appendix A.

2. Passage to the Hamilton–Jacobi equation in Minkowski space. The problem of finding all R-separable systems for (∗) is shown to be equivalent to that of finding all separable systems for the Helmholtz equation (2.1) defined on Minkowski space with symmetry (2.2). The conditions for separability of (2.1) are best derived from its classical counterpart, the Hamilton–Jacobi equation (2.15). We show that the symmetry algebras of (∗) and (2.15) are isomorphic to the Schrödinger algebra when $E=0$ and to the Galilean algebra when $E\neq 0$. The central result is the following theorem.

THEOREM 2.1. *There is a mapping between R-separable coordinate systems of* (∗) *and separable coordinate systems for the Helmholtz equation on* $n(=m+2)$ *dimensional Minkowski space.*

$$\square_n \bar{\Psi} = E\bar{\Psi}, \qquad n=m+2, \tag{2.1}$$

whose solutions $\bar{\Psi}$ *are eigenfunctions of the symmetry operator*

$$\partial_{\bar{x}^n} = \tfrac{1}{2}\left(\partial_{\bar{y}^{n-1}} + \partial_{\bar{y}^n}\right) \tag{2.2}$$

with associated eigenvalue ε (*i.e.* $\partial_{\bar{x}^n}\bar{\Psi} = \varepsilon\bar{\Psi}$). *Here the D'Alembertian operator on Minkowski space with the coordinates* $\bar{y}^i$, $i=1,2,\cdots,n$, *is defined by*

$$\square_n = \partial_{\bar{y}^1\bar{y}^1} + \cdots + \partial_{\bar{y}^{n-1}\bar{y}^{n-1}} - \partial_{\bar{y}^n\bar{y}^n}. \tag{2.3}$$

Proof. We first construct the mapping from (∗) to (2.1). Suppose

$$\begin{aligned} &y^j = y^j(x^1,\cdots,x^{n-1}), \qquad j=1,2,\cdots,n-2,\\ &t = t(x^1,\cdots,x^{n-1}), \end{aligned} \tag{2.4}$$

is an R-separable system for (∗); then there are functions Ψ, Ψ_j and R such that

$$\Psi = e^R \prod_{j=1}^{n-1} \Psi_j(x^j). \tag{2.5}$$

Consider the coordinate system $\bar{x}^i$

$$\begin{aligned} &\bar{y}^j = y^j(\bar{x}^1,\cdots,\bar{x}^{n-1}), \qquad j=1,2,\cdots,n-2,\\ &\bar{y}^{n-1} - \bar{y}^n = 2t(\bar{x}^1,\cdots,\bar{x}^{n-1}),\\ &\bar{y}^{n-1} + \bar{y}^n = \bar{x}^n - R/\varepsilon. \end{aligned} \tag{2.6}$$

This is a system in Minkowski space with ignorable $\bar{x}^n$, corresponding to the symmetry operator (2.2) whose eigenvalue we shall specify as ε. Let

$$\bar{\Psi} = e^{\varepsilon\bar{x}^n} \prod_{j=1}^{n-1} \Psi_j(\bar{x}^j) = e^{[\varepsilon(\bar{y}^{n-1}+\bar{y}^n)+R]} \prod_{j=1}^{n-1} \Psi_j(\bar{x}^j). \tag{2.7}$$

As Ψ satisfies (∗), $\bar{\Psi}$ is easily shown to satisfy (2.1). In other words (2.6) is a separable coordinate system for the Helmholtz equation (2.1). We now show that the mapping from (2.4) to (2.6) is onto by constructing its inverse. If $\{\bar{\mathbf{x}}\}$ is a separable system for (2.1) with symmetry operator (2.2), then

$$\begin{aligned} &\bar{y}^j = \bar{y}^j(\bar{x}^1,\cdots,\bar{x}^{n-1}), \qquad j=1,2,\cdots,n-2,\\ &\bar{y}^{n-1} - \bar{y}_n = 2t(\bar{x}^1,\cdots,\bar{x}^{n-1}),\\ &\bar{y}^{n-1} + \bar{y}^n = \bar{x}^n + f(\bar{x}^1,\cdots,\bar{x}^{n-1}) \end{aligned} \tag{2.8}$$

which if we let

$$f = -R/\varepsilon \tag{2.9}$$

is the image of (2.4), and the theorem is proved. Q.E.D.

In fact (2.1) separates in the same coordinate systems for each nonzero value of E. Care must be taken when $E=0$. When this occurs, (2.1) is the wave equation which in general separates in additional coordinate systems as is shown in Kalnins and Miller (1982b). However if we define

$$\tilde{\Psi} = e^{-Et/2\varepsilon}\Psi \tag{2.10}$$

then by substitution:

$\tilde{\Psi}$ is an R-separable solution for $(*)$ with $E=0$ iff Ψ is an R-separable solution for $(*)$.

The presence of the symmetry (2.2) has eliminated these additional systems. This means that a classification for nonzero E will also yield all possible systems when $E=0$. Again care must be taken as the symmetry group of the wave equation is larger: it admits additional conformal symmetries. In essence, some of the systems which are inequivalent for $(*)$ when $E\neq 0$ become equivalent under the action of the extra symmetries when $E=0$. We will return to this point later. All that we need to know at the moment is that everything can be obtained from the nonzero case.

One of the advantages of transforming our problem to the Helmholtz equation (2.1) is that to such an equation we can naturally associate a Riemannian geometry. Here (2.1) corresponds to the metric

$$ds^2 = (dy^1)^2 + \cdots + (dy^{n-1})^2 - (dy^n)^2 \tag{2.11}$$

which is the metric of a general n-dimensional Minkowski space $E(n,1)$. The classification of coordinates can now be framed as the classification of the metrics $ds^2 = g_{ij}dx^i dx^j$ arising from the fundamental metric (2.11) using the results of Riemannian geometry. Since the space is Minkowski, these results imply that all the components of the Riemann curvature tensor are identically zero, i.e.

$$R_{ijkl} \equiv 0, \qquad 1 \leqq i,j,k,l \leqq n. \tag{2.12}$$

For the definition of this tensor the reader is directed to Eisenhart (1949). Continuing our discussion in general let V_n be a (local) pseudo-Riemannian manifold specified by its metric

$$ds^2 = g_{ij}dx^i dx^j, \qquad g = \det(g_{ij}) \neq 0. \tag{2.13}$$

The Helmholtz equation for V_n is expressed in local coordinates by

$$\frac{1}{g^{1/2}}\,\partial_i\left(g^{1/2}g^{ij}\partial_j\Psi\right) = E\Psi \tag{2.14}$$

where E is a nonzero constant. Closely associated with the Helmholtz equation is the Hamilton–Jacobi equation

$$H(p_i) = g^{ij}p_i p_j = E, \quad p_i = \frac{\partial W}{\partial x^i}, \tag{2.15}$$

where W is the Hamilton–Jacobi function or complete integral of (2.15). This equation is said to be separable if there are functions W_i such that

$$W(\mathbf{x},\boldsymbol{\lambda})=\sum_{i=1}^{n} W_i(x^i,\boldsymbol{\lambda}), \tag{2.16}$$

where

$$\boldsymbol{\lambda}=(\lambda_1,\cdots,\lambda_n) \tag{2.17}$$

are the n constants of the motion necessary for a complete integral. (N.B. W is a complete integral if $\det(\partial W_i/\partial\lambda_j)\neq 0$). One of these constants can always be taken as E. Both (2.14) and (2.15) are coordinate independent and the passage between them is analogous to that between Classical and Quantum mechanics. A close relationship also exists in the case of variable separation. Every separable system for the Helmholtz equation also separates the Hamilton–Jacobi equation. This was demonstrated for orthogonal coordinates by Robertson (1927), and can be verified for nonorthogonal coordinate systems by using the results of Kalnins and Miller (1983). Robertson (1927) also supplied the extra condition (called the Robertson condition) for the converse of his result, i.e. for Hamilton–Jacobi separation to imply Helmholtz separation. The generalisation of the Robertson condition to nonorthogonal coordinates can be obtained from the work of Kalnins and Miller (1983).

The above results imply that every separable coordinate system of (2.1) is a separable system for the Hamilton–Jacobi equation

$$H=P_1^2+\cdots+P_{n-1}^2-P_n^2=E, \qquad P_i=\frac{\partial W}{\partial y^i} \tag{2.18}$$

with Minkowski space metric (2.11) and symmetry operator (2.2). The generalised Robertson condition depends on the metric components and is especially complicated in metrics with many nonorthogonal terms. In our case this condition can be shown to be always satisfied after the metric has been reduced to a simple form in §3. That is, the Hamilton–Jacobi equation and (2.1) separate in exactly the same coordinate systems. This result can also be proved by making the following observation which avoids introducing the technical details of the generalised Robertson condition. The separability of (2.1) in all separable systems for the Hamilton–Jacobi equation (2.18) is ultimately demonstrated in §7 by obtaining the corresponding separation equations for (*). In summary our problem has been reduced to the much easier one of finding all separable systems for the Hamilton–Jacobi equation (2.18).

Working with the Hamilton–Jacobi equation is more convenient since its form (2.15) is not as complicated as that of the Helmholtz equation (2.1) and is more closely related to the metric (2.13) (it is just the inverse). Using his definition of *equivalence*, Benenti (1980) was able to give the conditions for separability of the Hamilton–Jacobi equation (2.18). He showed that each of his classes contained a *canonical* separable system $\{x^a,x^r,x^\alpha\}$ with contravariant metric

$$(g^{ij})=\begin{pmatrix} \delta^{ab}g^{aa} & 0 & 0 \\ 0 & 0 & g^{r\alpha} \\ 0 & g^{r\alpha} & g^{\alpha\beta} \end{pmatrix}\begin{matrix} n_1 \\ n_2 \\ n_3 \end{matrix}. \tag{2.19}$$

Here $n_3 \geqq n_2$ and the integers a, r, α vary in the ranges $1 \leqq a,\ b \leqq n_1;\ n_1+1 \leqq r \leqq n_1+n_2;$ $n_1+n_2+1 \leqq \alpha,\ \beta \leqq n_1+n_2+n_3=n$. The nonzero components of the contravariant metric (2.19) are

$$g^{aa}=\frac{\psi^{a1}}{\psi},\quad g^{r\alpha}=B_r^{r\alpha}(x^r)\frac{\psi^{r1}}{\psi},\quad g^{\alpha\beta}=\sum_b A_b^{\alpha\beta}(x^b)\frac{\psi^{b1}}{\psi}, \tag{2.20}$$

where ψ and ψ^{i1} are the determinant and $i1$ cofactors of the $(n_1+n_2)\times(n_1+n_2)$ Stäckel matrix $(\psi_{ij}(x^i))$. Furthermore, $\partial_\alpha g^{ij}=0$ for each *ignorable* variable x^α. The variables x^a and x^r are referred to as *Stäckel* and *first order* variables respectively. Together these variables form the class of essential variables that was mentioned in the Introduction. The form of (2.19) implies that the first order variables are of null-type, and the signature of Minkowski space which is $n-1$ thus limits the number of these variables (n_2) to 1. A discussion of null variables is given in Eisenhart (1949).

Since our problem will be solved using the Hamilton–Jacobi equation (2.18), we first find the Killing vectors corresponding to the first order symmetries of (2.1) and (*) when $E \neq 0$. To find these Killing vectors, we solve the analogue of (1.5):

$$\{H, \mu\}_P = 0 \tag{2.21}$$

where $\{\ ,\ \}_P$ is the Poisson bracket and $\mu = a^j(\mathbf{y})P_j$. When $E \neq 0$ the Hamilton–Jacobi equation (2.18) admits a Lie algebra of Killing vectors which is the Lie algebra of Minkowski space $e(n,1)$. All the coordinate systems we are considering possess the symmetry (2.2) and so by the abelian subalgebra condition all symmetries of (2.1) must commute with (2.2). The Killing vector counterpart of this relation is

$$\{\hat{\varepsilon}, \mu\}_P = 0, \qquad \hat{\varepsilon} = \tfrac{1}{2}(P_{n-1}+P_n) \tag{2.22}$$

where $\hat{\varepsilon}$ is the Killing vector corresponding to (2.2). This extra condition confines us to a subalgebra of $e(n,1)$—the Galilean algebra g_m of dimension $\frac{1}{2}m(m+1)+2$ and basis

$$\hat{\varepsilon},\ P_u,\ M_{uv}=y^uP_v-y^vP_u,\quad B_u=y^u\hat{\varepsilon}-tP_u,\quad K_{-2}=P_t. \tag{2.23}$$

Here $u, v \in U=\{1,\cdots,n-2\}$. In general the indices u and v will be taken from the set U. Since the above Killing vectors form a basis, any Killing vector has the form

$$\lambda_\alpha = \rho_\alpha^u P_u + m_\alpha^{uv}M_{uv} + \beta_\alpha^u B_u + \kappa_\alpha K_{-2}, \tag{2.24}$$

where the ρ_α^u, m_α^{uv}, β_α^u and κ_α are all real constants. The commutation relations for g_m may be derived from Table 2.1. The symmetry algebra for the Helmholtz equation (2.1) ($E \neq 0$) is again g_m with the identifications

$$\begin{gathered}\hat{\varepsilon} \to \tfrac{1}{2}(\partial_{n-1}+\partial_n),\quad P_u \to \partial_u,\quad M_{uv} \to y^u\partial_v - y^v\partial_u,\\ B_u \to y^u\hat{\varepsilon} - t\partial_u,\qquad K_{-2} \to \partial_t.\end{gathered} \tag{2.25}$$

By solving (1.5) for (*) we find that g_m is also the symmetry algebra for (*) except that the operator $\hat{\varepsilon}$ is replaced by its eigenvalue ε.

If $E=0$ the symmetries of (2.1) are of the form a^jP_j+f where $\{H,a^jP_j+f\}_P=M(\mathbf{Y})H$ and then there are two extra symmetries.[1] When expressed as conformal Killing vectors, these are:

$$K_2=-t^2P_t-t\sum_u y^uP_u+\frac{1}{2}\hat{\varepsilon}\sum_u(y^u)^2, \qquad D=\sum_u y^uP_u+2tP_t, \tag{2.26}$$

and as operators for (*)

$$K_2=-t^2\partial_t-t\sum_u y^u\partial_u-\frac{1}{2}mt+\frac{1}{2}\varepsilon\sum_u(y^u)^2, \qquad D=\sum_u y^u\partial_u+2t\partial_t+\frac{1}{2}m. \tag{2.27}$$

These satisfy the commutation relations

$$\begin{aligned}&[D,P_u]=-P_u,\quad [D,M_{uv}]=0,\quad [D,B_u]=B_u,\quad [D,K_{\pm 2}]=\pm 2k_{\pm 2},\\ &[K_2,P_u]=-B_u,\quad [K_2,M_{uv}]=0,\quad [K_2,B_u]=0,\quad [K_2,K_{-2}]=D.\end{aligned} \tag{2.28}$$

This enlarged algebra is the Schrödinger algebra s_m of dimension $\frac{1}{2}m(m+3)+4$.

Geometrically the P_u and M_{uv} are generators of space translations and rotations under which (*) is clearly invariant. Less obvious are the transformations corresponding to the B_u's. These are the Galilean or velocity boosts and illustrate the fact that (*) retains its form in uniformly moving frames of reference. When $E=0$ we have the additional conformal Killing vectors D and K_2. D is the generator of the dilatation symmetry $\Psi(\mathbf{y},t)\to\Psi(\alpha\mathbf{y},\alpha^2t)$. The action of K_2 is rather complicated and it does not have a simple geometrical interpretation.

One of the applications of these symmetry algebras will be to determine the abelian subalgebras corresponding to sets of ignorable variables. Again the action of the group helps us choose simple representatives for these subalgebras. The group acts on the algebra g_m via its *adjoint action*. In general, for two members of a Lie algebra L_1 and L_2, the adjoint action of L_1 on L_2 is given by

$$e^{aL_1}L_2e^{-aL_1}=e^{a\,\mathrm{Ad}\,L_1}L_2 \tag{2.29}$$

where $\mathrm{Ad}\,L_1(L_2)\equiv[L_1,L_2]$. A proof of this result is given by Hausner and Schwartz (1968). The adjoint actions for the Galilean algebra are summarised in Table 2.1. In Table 2.1, each entry represents $e^{a\,\mathrm{Ad}\,L_1}(L_2)$, e.g. $e^{a\,\mathrm{Ad}\,P_u}(B_w)=B_w+a\delta_{uw}\varepsilon/2$. If the adjoint has no effect, i.e. $e^{a\,\mathrm{Ad}\,L_1}(L_2)=L_2$, then there is no entry. From (2.29) the commutation relations are simply the coefficients of a in the table. For example $[M_{uv},P_w]=\delta_{vw}P_u-\delta_{uw}P_v$ where δ_{jk} is the Kronecker delta.

The operators P_u, B_u, ε generate the $(2m+1)$-dimensional Weyl algebra w_m and the M_{uv}'s generate the $\frac{1}{2}m(m-1)$-dimensional orthogonal algebra $o(m)$. If we define

---

[1] The referee has commented that if we consider E-dependent operators, then

$$K_2=-t^2\left(\partial_t-\frac{E}{2\varepsilon}\right)-t\sum_u y^u\partial_u-\frac{1}{2}mt+\frac{1}{2}\varepsilon\sum_u(y^u)^2,$$

$$D=2t\left(\partial_t-\frac{E}{2\varepsilon}\right)+\sum_u y^u\partial_u+\frac{1}{2}m+2\varepsilon Et,$$

are conformal symmetries of (*) even when $E\neq 0$. This reflects the fact that solutions for $E\neq 0$ can be mapped to those for $E=0$ by the E-dependent relation (2.10).

the operators

$$A_1 = D, \quad A_2 = K_2 + K_{-2}, \quad A_3 = K_{-2} - K_2, \tag{2.30}$$

these satisfy the commutation relations

$$[A_1, A_2] = -2A_3, \quad [A_3, A_1] = 2A_2, \quad [A_2, A_3] = 2A_1 \tag{2.31}$$

and form a basis for the Lie algebra $sl(2,\mathbb{R})$. It follows that the structure of s_m is

$$s_m = (sl(2,\mathbb{R}) \oplus o(m)) \oplus w_m. \tag{2.32}$$

where $\oplus$ represents the direct sum and $\oplus$ the indirect sum. Similarly the Galilean algebra has structure

$$g_m = (t_1 \oplus o(m)) \oplus w_m \tag{2.33}$$

where t_1 is the one-dimensional algebra of time translations. Using standard results from Lie theory, these operators can be exponentiated to obtain the Schrödinger and Galilean groups. These groups act on the space of locally analytic functions of the real variables y^j, t and map solutions of (*) into solutions. Expressions for the actions of these groups appear in Miller (1977).

TABLE 2.1

Adjoint actions for the Galilean algebra.

L_1 \ L_2	P_w	M_{wz}	B_w	K_{-2}
P_u		$M_{wz} + a(\delta_{uw}P_z - \delta_{uz}P_w)$	$B_w + \frac{a\delta_{uw}\varepsilon}{2}$	
M_{uv}	$P_w + a(\delta_{vw}P_u - \delta_{uw}P_v) + a^2 \ldots$ †	$M_{wz} + a(\delta_{vw}M_{uz} - \delta_{vz}M_{uw} + \delta_{uz}M_{vw} - \delta_{uw}M_{vz}) + a^2 \ldots$ †	$B_w + a(\delta_{vw}B_u - \delta_{uw}B_v) + a^2 \ldots$ †	
B_u	$P_w - \frac{a\delta_{uw}\varepsilon}{2}$	$M_{wz} + a(\delta_{uw}B_z - \delta_{uz}B_w)$		$K_{-2} + aP_u - \frac{a^2\varepsilon}{4}$
K_{-2}			$B_w - aP_w$	

†The adjoint action of M_{uv} on the P_w's is $e^{a\,\mathrm{Ad}\,M_{uv}}(\rho_u P_u + \rho_v P_v) = \rho'_u P_u + \rho'_v P_v$ where ρ'_u and ρ'_v are determined by

$$\begin{pmatrix} \rho'_u \\ \rho'_v \end{pmatrix} = \begin{pmatrix} \cos(a) & \sin(a) \\ -\sin(a) & \cos(a) \end{pmatrix} \begin{pmatrix} \rho_u \\ \rho_v \end{pmatrix}$$

It acts on the B_u's in exactly the same manner. $(u, v, w, z \in U = \{1, \cdots, n-2\})$.

3. Reducing the Hamilton–Jacobi equation. In this section we find simple forms for both the metric and the Killing vectors, summarising these results in the following theorem.

THEOREM 3.1. *All Hamilton–Jacobi separable coordinate systems for* (2.18) *with Killing vector* $\hat{\varepsilon}$, *are equivalent to a coordinate system associated with the Hamilton–Jacobian equation*

$$g^{11}p_1^2 + \cdots + g^{(n-2)(n-2)}p_{n-2}^2 + g^{nn}p_n^2 + 2p_{n-1}p_n = E. \tag{3.1}$$

Moreover if one variable is first order, and there are r ignorables $x^{\alpha(1)},\cdots,x^{\alpha(r)}$ in addition to x^n, the corresponding Killing vectors are

$$
\begin{aligned}
&\lambda_n=\frac{1}{2}(P_{n-1}+P_n)=\hat{\varepsilon},\\
&\lambda_{\alpha(1)}=M_{12},\cdots,\lambda_{\alpha(p)}=M_{2p-1,2p},\\
&\lambda_{\alpha(p+1)}=P_{2p+1},\cdots,\lambda_{\alpha(q)}=P_{p+q},\\
&\lambda_{\alpha(q+1)}=B_{p+q+1}-\rho^{p+q+1}P_{p+q+1},\cdots,\lambda_{\alpha(r)}=B_{p+r}-\rho^{p+r}P_{p+r}.
\end{aligned}
\tag{3.2}
$$

If there are no first order variables then all the Killing vectors except for one[2] will be those in (3.2).

Proof. Let x^a be a Stäckel variable while x^r is first order. From (2.8)

$$g_{ni}=\frac{\partial t}{\partial x^i} \tag{3.3}$$

so that

$$\frac{\partial t}{\partial x^a}=0,\qquad \frac{\partial t}{\partial x^\alpha}=g_{n\alpha}(x^r), \tag{3.4}$$

since inversion of (2.19) shows that for Stäckel variables $g_{ai}=\delta^{ai}(g^{aa})^{-1}$. The equations in (3.4) can be integrated to give

$$t=h(x^r)+\sum_\alpha g_{n\alpha}(x^r)x^\alpha. \tag{3.5}$$

Now from (3.5) and (2.8)

$$\lambda_\alpha=y^u_\alpha P_u+g_{n\alpha}(P_{n-1}-P_n)+\tfrac{1}{2}f,_\alpha(P_{n-1}+P_n),\quad \alpha\neq n, \tag{3.6}$$

so (2.24) implies

$$g_{n\alpha}=\kappa_\alpha. \tag{3.7}$$

We know from §2 that $n_2\leqq 1$. If $n_2=0$ then;

$$t=\kappa_\alpha x^\alpha\to x^{n-1} \tag{3.8}$$

under a general linear motion amongst the ignorables. Consider $n_2=1$. If for some β, $\kappa_\beta\neq 0$, we can use a general linear motion and then the freedom (1.8b) to transform t to x^{n-2}, i.e.,

$$t\to h(x^{n-1})+x^{n-2}\to x^{n-2}. \tag{3.9}$$

Thus from (3.7), the metric g_{ij} has form:

$$ds^2=\sum_{a=1}^{n_1} g_{aa}(dx^a)^2+\sum_{k,l} g_{kl}dx^k dx^l+2dx^{n-2}dx^n \qquad (n_1+1\leqq k,l\leqq n-1). \tag{3.10a}$$

---

[2] This Killing vector can only be calculated from the coordinate transformations (4.30). It has form $K_{-2}+\beta^u B_u$, but its knowledge was not needed to produce the reduced form of the contravariant metric (3.1). It may be regarded as a special subcase when the first order variable has become ignorable.

The variable x^{n-1} is first order, thus $g^{(n-1)(n-1)}=0$ and by taking the inverse of (3.10a), (g^{ij}) has form

(3.10b)
$$\begin{pmatrix} * & & & & & & & & \\ & \ddots & & & & & & & \\ & & * & & & & & & \\ & & & * & * & \cdots & * & 0 & * & * \\ & & & * & & & \cdot & 0 & * & \cdot \\ & & & \cdot & & & \cdot & \cdot & \cdot & \cdot \\ & & & * & \cdot & \cdots & * & \cdot & \cdot & * \\ & & & 0 & \cdot & & \cdot & 0 & 0 & 1 \\ & & & * & * & \cdots & * & 0 & 0 & * \\ & & & * & \cdot & \cdots & * & 1 & * & * \end{pmatrix}$$

The vectors

(3.11)
$$\begin{aligned} \mathbf{X}^{n-1} &= (0,0,\cdots,0,0,1,0), \\ \mathbf{X}^{n-2} &= (0,0,\cdots,0,1,0,0) \end{aligned}$$

are orthogonal and null with respect to (g^{ij}), which is impossible in Minkowski space (see Eisenhart (1949)). Thus

(3.12)
$$\kappa_\alpha = 0, \quad \text{for } \textit{each } \alpha.$$

It is now possible to use the freedom (1.8a) to change variables so that

(3.13)
$$h(x^{n-1}) \to x^{n-1}.$$

We see that whether x^{n-1} is first order or ignorable the coordinate transformations are

(3.14)
$$\begin{aligned} y^u &= y^u(x^1,\cdots,x^{n-1}), \\ y^{n-1} - y^n &= 2x^{n-1} = 2t, \\ y^{n-1} + y^n &= x^n + f(x^1,\cdots,x^{n-1}). \end{aligned}$$

This is progress as the time coordinate t has been identified as just x^{n-1}.

From (2.24) and (3.14)

(3.15)
$$\lambda_\alpha = \rho_\alpha^u P_u + m_\alpha^{uv} M_{uv} + \beta_\alpha^u B_u, \qquad \alpha \neq n-1.$$

To solve the orbit problem, i.e., to establish the possible commuting sets of Killing vectors, we refer the reader to Reid (1984). The essential element of that argument is to notice that when we set time to be a constant, i.e., $x^{n-1} = \textit{constant} = c$, in the coordinate transformations (3.14), then we have a separable system for $\Delta_m \Psi = E\Psi$, parameterised by c. The results for the orbit analysis for this equation are known (see Kalnins and Miller (1982a)), and are used in Reid (1984) to determine some of the unknown constants in (3.15). Further simplification of the orbits is achieved through the use of the adjoint actions given in Table 2.1. After some work the Killing vectors are shown to have the form given in (3.2). Solving the characteristic equations resulting from (3.2) shows that the metric can be written as

(3.16)
$$ds^2 = \sum_u g_{uu}(dx^u)^2 + g_{(n-1)(n-1)}(dx^{n-1})^2 + 2\,dx^{n-1}\,dx^n.$$

Inversion of (3.16) gives the Hamilton–Jacobi equation in the form (3.1)—all the conditions of Theorem 3.1 have now been satisfied.

By finding the symmetries and using the restrictive signature of Minkowski space we have reduced the nonorthogonal part of the Hamilton–Jacobi equation to one off-diagonal element.

4. Metric and coordinates. In the previous section we established a simple form (3.1) for the contravariant metric and this will enable us to use the curvature conditions (2.12) to determine this form exactly. We will also find the coordinate transformations.

THEOREM 4.1. *The Hamilton–Jacobi equation can be written*

$$\sum_{q\in Q}\frac{1}{\sigma_q}\sum_{b\in B_q}\bar{g}^{bb}p_b^2+2p_{n-1}p_n+\sum_{q\in Q}\frac{V_q}{\sigma_q}p_n^2=E, \tag{4.1}$$

where

$$Q=\{q_i\colon q_1<q_2<\cdots<q_l\} \tag{4.2}$$

is some subset of $U=\{1,\cdots,m\}$ *and the sets*

$$B_{q_i}=\{q_i,q_i+1,\cdots,q_{i+1}-1\},\qquad 1\leqq i\leqq l, \tag{4.3}$$

form a partition of U. *Here* σ_q *is a function of* x^{n-1} *alone, and if* $b\in B_q$

$$\bar{g}^{bb}=\bar{g}^{bb}(x^c),\qquad V_q=V_q(x^c) \tag{4.4}$$

where $c\in B_q$.

As an illustration of the notation used in Theorem 4.1 see (5.14). There $U=Q=\{1,2\}$, $B_1=\{1\}$ and $B_2=\{2\}$. In (5.17) $Q=\{1\}$, $B_1=\{1,2\}$ and $\sigma_1=|(x^3)^2+1|^{1/2}$. To prove the Theorem, we will make use of the following equivalent condition for the Stäckel matrix found by Eisenhart (1934).

LEMMA 4.1. *The nonsingular* $Q\times Q$ *matrix* (ψ_{ij}) *is a Stäckel matrix if and only if*

$$\begin{aligned}\partial_{jk}\log\left(\frac{\psi}{\psi^{i1}}\right)&=\partial_j\log\left(\frac{\psi}{\psi^{i1}}\right)\partial_k\log\left(\frac{\psi}{\psi^{i1}}\right)-\partial_j\log\left(\frac{\psi}{\psi^{i1}}\right)\partial_k\log\left(\frac{\psi}{\psi^{j1}}\right)\\&\quad-\partial_k\log\left(\frac{\psi}{\psi^{i1}}\right)\partial_j\log\left(\frac{\psi}{\psi^{k1}}\right),\qquad 1\leqq j<k\leqq Q.\end{aligned} \tag{4.5}$$

We now prove Theorem 4.1.

Proof of Theorem 4.1. For notational convenience we need only assume $n_2=1$ since the functional form for $n_2=0$ is just a special case. Consider the matrix

$$(\Phi_{ij})=\begin{pmatrix}\psi_{ab} & 0\\ A_{\beta a} & I_{MM}\end{pmatrix} \tag{4.6}$$

where

$$A_{\beta a}=-A_a^{\beta\beta},\qquad \beta=n_1+1,\cdots,n-2 \tag{4.7}$$

is an $M\times N$ matrix ($N=n_1+1, M=n-2-n_1$) and I_{MM} is the $M\times M$ identity matrix. This matrix has the properties

$$g^{aa}=\frac{\Phi^{a1}}{\Phi},\quad g^{\beta\beta}=\frac{\Phi^{\beta 1}}{\Phi},\quad g^{(n-1)n}=1=\frac{\Phi^{(n-1)1}}{\Phi}. \tag{4.8}$$

These will enable us to make extensive use of Lemma 4.1. It follows immediately from (3.1) and Lemma 4.1 that

$$\partial_{v(n-1)}\log g_{uu}=\partial_v\log g_{uu}\partial_{n-1}\log g_{uu}-\partial_v\log g_{uu}\partial_{n-1}\log g_{vv}. \tag{4.9}$$

It is easily shown that

$$R_{vuu(n-1)}=\frac{1}{2}g_{uu}\partial_{v(n-1)}\log(g_{uu}) + \frac{g_{uu}}{4}\left[\partial_v\log(g_{uu})\partial_{n-1}\log g_{uu}-\partial_v\log(g_{uu})\partial_{n-1}\log(g_{vv})\right]=0. \tag{4.10}$$

Substitution of (4.9) into $[\cdot]$ in (4.10) gives

$$\partial_{v(n-1)}\log(g_{uu})=0. \tag{4.11}$$

This implies

$$g_{uu}=\sigma_u(x^{n-1})\bar{g}_{uu} \tag{4.12}$$

where $\bar{g}_{uu}$ does not depend on x^{n-1}. From (4.9) and (4.11)

$$\partial_v\log(g_{uu})\,\partial_{n-1}\log(\sigma_u/\sigma_v)=0. \tag{4.13}$$

We can define an equivalence relation $\sim$ on U by $u\sim v$ if σ_u is proportional to σ_v, and by rescaling coordinates there is no loss in assuming that $\sigma_u=\sigma_v$ on each equivalence class. By reordering the indices, the sets of the partition can be taken to be the B_q's of Theorem 4.1. Furthermore, (4.13) implies that the $\bar{g}_{bb}$'s are only functions of those x^c's in their class, i.e. they satisfy property (4.4) in Theorem 4.1. Defining

$$V_q=\sum_{b\in B_q}\bar{g}^{bb}A_b^{nn}, \tag{4.14}$$

the Hamilton–Jacobi equation takes the form (4.1) given in Theorem 4.1. Q.E.D.

THEOREM 4.2. *The spaces with metrics*

$$d\bar{s}_q^2=\sum_{b\in B_q}\bar{g}_{bb}(dx^b)^2 \tag{4.15}$$

are separable, flat and positive definite. Also we have

$$\partial_{x^cx^d}V_q=0,\quad c\neq d,\quad c,\,d\in B_q. \tag{4.16}$$

Proof. We first show that the metrics (4.15) are differentially flat, separable and positive definite. The flatness conditions $R_{ijkl}=0$ with $i,j,k,l\in B_q$, are equivalent to $\bar{R}_{ijkl}=0$, i.e., those for the metrics in (4.15). It follows that the spaces associated with these metrics are flat. Since the metrics in (4.15) are orthogonal, the separability conditions (4.5) for $i,j,k\in B_q$ can be applied to show that these metrics are also separable. To show that the metrics in (4.15) are positive definite, we first compute the eigenvalues of (4.1). These are

$$\xi_u=g^{uu},\qquad \xi_\pm=\tfrac{1}{2}\left[g^{nn}\pm\sqrt{(g^{nn})^2+4}\right]. \tag{4.17}$$

Regardless of the value of g^{nn} the eigenvalues ξ_+ and ξ_- are positive and negative

respectively. The remaining eigenvalues ξ_u must be positive since the space is Minkowski with signature $n-1$. It now follows from (4.1) that

$$\sigma_q > 0 \quad \text{and} \quad \bar{g}_{uu} > 0, \tag{4.18}$$

redefining σ_q and $\bar{g}_{uu}$ to be their negatives if necessary. This shows that the metrics (4.15) are positive definite.

To prove condition (4.16) of the theorem we notice that since V_q is given by (4.14), the extension of Lemma 4.1 (see (4.8)) can be used to obtain

$$\partial_{cd}V_q + \partial_c V_q \partial_d \log(g_{cc}) + \partial_d V_q \partial_c \log(g_{dd}) = 0. \tag{4.19}$$

Alternatively this result can be obtained by considering the Hamilton–Jacobi equation (artificially created for our purpose) $\Sigma_{b\in B_q}\bar{g}^{bb}p_b^2 + \Sigma_q V_q p_{n-1}^2$. As this equation is separable the results of Lemma 4.1 can be applied: equation (4.19) is derived from (4.5) with $j=c$, $k=d$ and $i=n-1$. Combining this result with the curvature conditions $R_{c(n-1)(n-1)d}=0$, which are equivalent to

$$-\tfrac{1}{2}\partial_{cd}V_q + \partial_c V_q \partial_d \log(g_{cc})/4 + \partial_d V_q \partial_c \log(g_{dd})/4 = 0, \tag{4.20}$$

we obtain condition (4.16). Q.E.D.

We now go on to determine the exact form of the unknown functions $\bar{g}_{bb}$, σ_q, V_q, and hence obtain the classification we are seeking. First since the spaces defined by the metrics (4.15) are positive definite, we can use the results of Kalnins and Miller (1982a), in which they completely classified such spaces; this determines the $\bar{g}_{uu}$'s. By transforming to standard cartesian coordinates on each $\mathbb{R}^{n_q}$, the σ_q and V_q are determined. In transforming back to general separable coordinates, however, the separability is not necessarily preserved, so we find the compatibility conditions to ensure this preservation.

Since the spaces defined by the metrics are flat and positive definite we can choose standard coordinates z^b:

$$z^b = z^b(x^c), \qquad b, c \in B_q, \tag{4.21}$$

and

$$d\bar{s}_q^2 = \sum_{b\in B_q} (dz^b)^2. \tag{4.22}$$

Working in terms of these coordinates (4.19) implies that

$$\partial_{z^c z^d}V_q = 0, \quad c, d \in B_q, \quad c \neq d. \tag{4.23}$$

$R_{(n-1)cc(n-1)}=0$ is equivalent to

$$\frac{1}{2}\sigma_q'' - \frac{(\sigma_q')^2}{4\sigma_q} - \frac{1}{2}\sum_{t\in Q}\frac{V_{t,cc}}{\sigma_t} = 0 \tag{4.24}$$

where the prime denotes differentiation with respect to x^{n-1}. Together with (4.23) this last equation implies that

$$V_q = \sum_{b\in B_q}\left(\frac{\xi_q}{4}(z^b)^2 + \gamma_b z^b\right) + \delta \tag{4.25}$$

and

$$2\sigma_q\sigma_q''-(\sigma_q')^2=\zeta_q,\quad \sigma_q>0,\quad \zeta_q,\delta\in\mathbb{R}. \tag{4.26}$$

Making a transformation of form $x^n\to x^n+g(x^{n-1})$, we can take $\delta=0$ in (4.25). When $\zeta_q\neq 0$ it is possible to translate z^b so that $\gamma_b\to 0$. To solve (4.26), differentiate it to obtain $\sigma'''=0$ and substitute the resulting quadratic into the original equation (4.26). Application of the coordinate freedoms (1.8) leads to the five possibilities $\mathrm{I}\to\mathrm{IV}\pm$ in Table 4.1. The constants ζ_q, γ_b, v_q and w_q are all real.

TABLE 4.1

Possibilities for $\sigma(x^{n-1})$.

Type	σ_q	ζ_q	γ_q
I	1	0	arbitrary
II	$(x^{n-1}+v_q)^2$	0	arbitrary
III	$(x^{n-1}+v_q)$	-1	0
IV$\pm$	$\left\lvert(x^{n-1}+v_q)^2\pm w_q^2\right\rvert$	$\pm 4w_q^2$, $w_q\neq 0$	0

A knowledge of the lower dimensional cases, and a little guesswork leads to the transformations for $n=3(m=1)$. They are

$$\begin{aligned} &y^1=z^1\sigma^{1/2}+\frac{1}{2}\gamma\int\int\sigma^{-3/2},\\ &y^2+y^3=x^3-\sigma'(z^1)^2/4-\frac{1}{2}\gamma z^1\sigma^{1/2}\int\sigma^{-3/2}-\frac{\gamma^2}{8}\int\left(\int\sigma^{-3/2}\right)^2,\\ &y^2-y^3=2x^2, \end{aligned} \tag{4.27}$$

where

$$\int f(y)=\int_{y=y_0}^{x^2} f(y)\,dy. \tag{4.28}$$

We now show that this one-dimensional case forms the building block for all others. Define

$$F_{uq}(z^u,x^{n-1})=z^u\sigma_q^{1/2}+\frac{1}{2}\gamma_u\iint\sigma_q^{-3/2}, \tag{4.29}$$

and

$$G_{uq}(z^u,x^{n-1})=\sigma_q'(z^u)^2/4+\frac{1}{2}\gamma_u z^u\sigma_q^{1/2}\int\sigma_q^{-3/2}+\frac{\gamma_u^2}{8}\int\left(\int\sigma_q^{-3/2}\right)^2, \tag{4.30}$$

then the coordinate transformations are

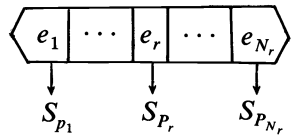
$$\begin{aligned} &y^u=F_{uq},\\ &y^{n-1}+y^n=x^n-\sum_{q\in Q}\sum_{b\in b_q}G_{bq},\\ &y^{n-1}-y^n=2x^{n-1}. \end{aligned} \tag{4.31}$$

These results can be checked by computing the metric using (4.31) and then taking its inverse to obtain the Hamilton–Jacobi equation (4.1). In the above transformations we will say that the γ_u term is *attached* to the z^u coordinate.

To obtain the coordinate transformations in terms of the x^i we simply substitute the expressions given for the z^b in (4.21). Not all the systems thus obtained will be separable. Naturally they will be separable when $\gamma_u=0$ for all u and $\zeta_q=0$, or for all values of these parameters when the coordinates are cartesian. Given a certain separable system (4.21) the problem is to work out the form of V_q (or equivalently the values of the parameters γ_u and ζ_q) to ensure separation in the variables x^i.

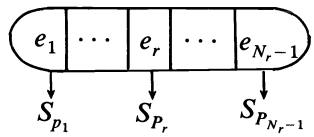
To tackle this problem it is necessary to know just what the possible systems (4.21) are. We give a brief summary of the solution provided by Kalnins and Miller (1982a). They show that it is possible to decompose $\mathbb{R}^n$ into a direct sum of subspaces $\mathbb{R}^{n_r}$ in such a way that the separable coordinates on each of these are of either "elliptic" or "parabolic" type with graphical representations

(4.32A)



or

(4.32B)



respectively. We will say that $\mathbb{R}^n$ *splits* into the subspaces $\mathbb{R}^{n_r}$. Cartesian coordinates on $\mathbb{R}^{n_r}$ for case A are given by

$$z^i=({}_{N_r}w_j)({}_{p_j}s_{q_j}),\qquad 1\leqq i\leqq n_r,\quad 1\leqq j\leqq N_r \tag{4.33A}$$

and for case B

$$\begin{aligned} &z^1=({}_{N_r}w_1),\\ &z^i=({}_{N_r}w_j)({}_{p_j}s_{q_j}),\qquad 2\leqq i\leqq n_r,\quad 2\leqq j\leqq N_r. \end{aligned} \tag{4.33B}$$

Here the ${}_{p_j}s_{q_j}$, $1\leqq q_j\leqq p_j=1$ are coordinates on the p_j-dimensional sphere S_{p_j} and therefore satisfy

$$\sum_{q_j=1}^{p_j+1} {}_{p_j}s_{q_j}^2=1. \tag{4.34}$$

In (4.32) the sphere S_{p_j} is said to be *attached* to e_j or equivalently to the coordinate ${}_{N_r}w_j$. When $p_j=0$, S_{p_j} is the "zero dimensional" or "trivial" sphere. Equivalently there is no sphere attached to ${}_{N_r}w_j$.

For elliptic-type coordinates A

$${}_{N_r}w_j^2=c_r^2\frac{\prod_{l=1}^{N_r}(x_r^l-e_j^r)}{\prod_{l\neq j}(e_l^r-e_j^r)},\qquad j=1,\cdots,N_r, \tag{4.35A}$$

where

$$e_1^r<x_r^1<\cdots<e_{N_r}^r<x_r^{N_r},\qquad c_r,\ e_j^r\in\mathbb{R}.$$

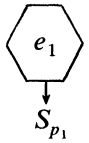
For parabolic coordinates B

$$
\begin{aligned}
{}_{N_r}w_1 &= \frac{c_r}{2}\left(\sum_{l=1}^{N_r} x_r^l + \sum_{l=1}^{N_r-1} e_l^r\right), \\
{}_{N_r}w_j^2 &= -c_r^2 \frac{\prod_{l=1}^{N_r}\left(x_r^l - e_{j-1}^r\right)}{\prod_{l\neq j-1}\left(e_l^r - e_{j-1}^r\right)}, \qquad j=2,\cdots,N_r,
\end{aligned}
\tag{4.35B}
$$

where

$$x_r^1 < e_1^r < \cdots < e_{N_r-1}^r < x_r^{N_r}, \qquad c_r,\ e_j^r \in \mathbb{R}.$$

If the case $N_r=1$ is treated in the same way as $N_r>1$, it is possible to have the elliptic and parabolic systems

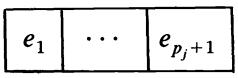


and

corresponding to the metrics $ds^2=(dx^1)^2/(x^1-e_1)$ and $ds^2=(dx^1)^2$ respectively. Both of these systems are equivalent via the scaling transformation (1.8a). This explains why the parabolic case does not appear when $N_r=1$. This will help clarify some of the exceptional behaviour that we will later encounter for $N_r=1$.

The classification also depends on the possible separable systems on the spheres S_{p_j}. The structure of these systems is independent of the V_q and σ_q terms. Separable systems on S_{p_j} can be built from irreducible blocks, each having graphical representation

(4.36)

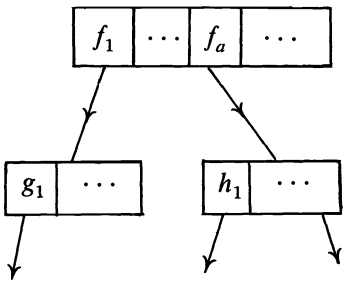


The coordinate transformations for this block are

$$
{}_p s_i^2 = c^2 \frac{\prod_{j=1}^{p}\left(x^j - e_i\right)}{\prod_{j\neq i}\left(e_j - e_i\right)}, \qquad 1 \leqq i \leqq p+1
\tag{4.37}
$$

where $e_1 < x^1 < \cdots < x^p < e_{p+1}$. A general graph for S_{p_j} has form

(4.38)



The coordinates for (4.38) can be built in multiplicative-iterative fashion from those of (4.37). This process is best understood by consulting Kalnins and Miller (1982a) and the examples provided in Tables 2 and 5 of the Appendix.

Applying these results to our case, decomposing the space corresponding to $d\bar{s}_q^2$, corresponds to partitioning B_q into subsets E_r in just the same way that we partitioned

U into the subsets B_q (i.e. the set E_r is labelled by its minimum value r). For example in (5.17), $E_1=\{1\}$, $E_2=\{2\}$, and in (5.22) $E_1=\{1,2\}$. Using this partition,

$$d\bar{s}_q^2=\sum_{r\in\mathscr{R}\cap B_q} d\bar{s}_r^2 \tag{4.39}$$

where $d\bar{s}_r^2$ is the infinitesimal distance

$$d\bar{s}_r^2=\sum_{b\in E_r}\bar{g}_{bb}(d\bar{x}^b)^2. \tag{4.40}$$

Here $\mathscr{R}$ is the set of all the (minimum) indices r (e.g. in both (5.14) and (5.17) $\mathscr{R}=\{1,2\}$). For simplicity on each irreducible block E_r of form (4.32a) or (4.32b) the coordinates x^r, $x^{r+1},\cdots,x^{r+n_r-1}$ are relabelled as x^1, $x^2,\cdots,x^{n_r}$. V_r for each of these blocks is defined as in (4.14) but with b restriced to E_r. Confining ourselves to one of these blocks E_r we systematically determine V_r.

Suppose that $\zeta_q\neq 0$ (i.e. σ_q is of types III or IV$\pm$) on E_r. Here

$$V_r=\sum_{i=1}^{n_r}\frac{\zeta_q}{4}(z^i)^2 \tag{4.41}$$

and from (4.33) and (4.34).

$$V_r=\frac{\zeta_q}{4}\sum_{j=1}^{N_r}{}_{N_r}w_j^2. \tag{4.42}$$

Substituting for the ${}_{N_r}w_j$ from (4.35) and expanding in partial fractions in terms of the e_j's, we find after some time that for the elliptic case A

$$V_r=\frac{\zeta_q c_r^2}{4}\left[\sum_{i=1}^{N_r}x^i-\sum_{i=1}^{N_r}e_i\right], \tag{4.43A}$$

and for the parabolic case B

$$V_r=\frac{\zeta_q c_r^2}{16}\left[2\sum_{i=1}^{N_r}(x^i)^2-\left(\sum_{i=1}^{N_r}x^i\right)^2+6\sum_{i=1}^{N_r}x^i\sum_{t=1}^{N_r-1}e_t-\left(\sum_{t=1}^{N_r-1}e_t\right)^2-2\sum_{t=1}^{N_r-1}e_t^2\right]. \tag{4.43B}$$

However only the V_r of (4.43A) satisfies (4.16). This means that if $\zeta_q\neq 0$ only elliptical-type coordinates are separable.

Now consider $\zeta_q=0$ and suppose that at least one of the γ_i's (γ_k say) is nonzero. In this case σ_q is of type I or II and

$$V_r=\sum_1^{n_r}\gamma_i z^i=\sum_{j=1}^{N_r}{}_{N_r}w_j\sum_{q_j=1}^{p_j+1}\gamma^i_{q_j p_j}s_{q_j}. \tag{4.44}$$

Let x^i be any of the separable coordinates on the sphere S_{p_1}. From (4.16)

$$\partial_{x^1x^i}V_r\left(x^1,x^2=e_2,\cdots,x^{N_r}=e_{N_r}\right)=0=\partial_{x^1}({}_{N_r}w_1)\,\partial_{x^i}\left(\sum_{q_1=1}^{p_1+1}\gamma^1_{q_1 p_1}s_{q_1}\right) \tag{4.45}$$

since ${}_{N_r}w_k(x^k=e_k)=0$. Now $\partial_{x^l}({}_{N_r}w_1)\neq 0$ so that

$$\partial_{x^l}\left(\sum_{q_1=1}^{p_1+1}\gamma^1_{q_1 p_1}{}_{p_1}s_{q_1}\right)=0. \tag{4.46}$$

i.e. $\sum_{q_1=1}^{p_1+1}\gamma^1_{q_1 p_1}{}_{p_1}s_{q_1}=A$, a constant. Such a relation cannot exist among the ${}_{p_1}s_{q_1}$'s unless $p_1=0$. No sphere can be attached to the coordinate ${}_{N_r}w_1$. This is demonstrated as follows. As the $\gamma^1_{q_1}$'s are real, rotations can be used to pass to an equivalent set of separable coordinates ${}_{p_1}\tilde{s}_{q_1}$, for the sphere S_{p_1} such that

$$\sum_{q_1=1}^{p_1+1}\gamma^1_{q_1 p_1}{}_{p_1}s_{q_1}\to\left[\sum_{q_1}\left(\gamma^1_{q_1}\right)^2\right]^{1/2}_{p_1}\tilde{s}_1. \tag{4.47}$$

Since $\sum_{q_1}(\gamma^1_{q_1})^2\neq 0$ equation (4.46) now implies

$$\partial_{x^l}({}_{p_1}\tilde{s}_1)=0, \quad \text{for } \textit{all } l. \tag{4.48}$$

This is only possible if $p_1=0$, i.e, no sphere is attached to the ${}_{N_r}w_1$ coordinate.

Still assuming $\zeta_q=0$ and $\gamma_k\neq 0$, consider the elliptic case. Let $N_r=1$. No spheres are attached to ${}_1w_1$ by the argument above and so $z^1={}_1w_1=x^1$. Condition (4.16) is satisfied and the system is separable. Let $N_r\geqq 2$. Using the above argument and the symmetry of the elliptic coordinate transformations (4.35A) shows that no spheres (beside the trivial ones) can be attached to any of the ${}_{N_r}w_j$ coordinates. Let x^1 and x^2 be any two of the elliptic-type coordinates in (4.35A). Equation (4.16) yields

$$\partial_{x^1x^2}V_r=0=\sum_i\frac{\gamma_i\,{}_{N_r}w_i}{4(x^1-e_i)(x^2-e_i)}. \tag{4.49}$$

Multiplying this last expression by $(x^2-e_2)^{1/2}$ and setting $x_i=e_i$ for $i\geqq 2$, we find that $\gamma_2=0$. This result is easily generalised to show that $\gamma_i=0$, for all i: contradicting our initial assumption that one of the γ_i's was nonzero. Under these conditions, it follows that the elliptic block E_r leads to separability only if $V_r=0$.

In the parabolic case using the same methods we find that only γ_1 can be nonzero and then

$$V_r=\gamma_1\,{}_{N_r}w_1=\frac{1}{2}\gamma_1 c_r\left(x^1+\cdots+x^{N_r}+e_1+\cdots+e_{N_r-1}\right). \tag{4.50}$$

i.e. parabolic coordinates are separable in this case.

Finally if $\gamma_1=0=\zeta_q$ (i.e. σ_q is of type I or II), then both elliptic or parabolic type blocks are possible.

We have determined all Euclidean coordinate systems that can combine with a given V_q to form a separable system for (4.1). Thus using the results of §2 the R-separation problem for (*) has been solved. This procedure is systematised in graphical form in the next section.

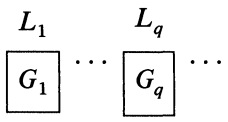
5. Graphical representation of coordinates. We develop a graphical calculus to represent the R-separable coordinate systems for (*), illustrating this procedure in detail for $m=1,2,3$—the cases of physical interest.

From (4.28) and (4.30) the coordinate transformations for (*) are given by

$$y^u=z^u\sigma_q^{1/2}+\frac{1}{2}\gamma_u\iint\sigma_q^{-3/2}, \qquad u\in B_q, \quad q\in Q, \tag{5.1}$$
$$t=x^{n-1},$$

and these can be given the graphical representation

(5.2)



Here G_q is a separable system on $\mathbb{R}^{n(B_q)}$ and the box indicates that there is just one function σ_q on G_q, while the Latin number L_q(I, II, III or IV$\pm$) specifies its type. G_q is one of Kalnins and Miller's graphs and will in general have form

$$G_q^1 \cdots G_q^r \cdots \tag{5.3}$$

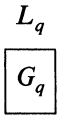
where each G_q^r is one of the elliptic or parabolic types given in (4.32A) and (4.32B).

At the end of the last section we found the compatibility conditions for a $\sigma_q - G_q^r$ combination to be separable. To summarise these results we start with a given separable Euclidean system G_q^r and then give the compatible σ functions.

If G_q^r is of elliptic-type, then L_q could take any of its values I $\to$ IV$\pm$. However the cases $\zeta_q = 0$, $\gamma_u \neq 0$ where x^u is one of the coordinates on G_q^r can only occur if G_q^r has form $\langle\!\overline{e}\!\rangle$. If the block G_q^r is of parabolic-type, it is only compatible with the σ-types I and II since it was shown in §4 that $\zeta_q \neq 0$ did not satisfy (4.16). In the allowed cases I and II, the γ_u term *attaches* itself to the coordinate ${}_{N_r}w_1$ of (4.35B).

These results are now generalised to a block of form

(5.4)



If all the G_q^r's are of elliptic-type, then L_q can take any of its values, I $\to$ IV$\pm$. The parameters γ_u on G_q can only be nonzero if the blocks G_q^r to which they are attached have form $\langle\!\overline{e}\!\rangle$.

If at least one G_q^r is of parabolic-type, then L_q is restricted to types I or II. The exceptional case $N = 1$ can be given our uniform general treatment if we regard it as being equivalent to the two systems $\bigcirc$ and $\langle\!\overline{e}\!\rangle$ as discussed in §4. A γ_u term can be attached to $\bigcirc$ (no spheres can be attached to this graph). Spheres can be attached to $\langle\!\overline{e}\!\rangle$ and this is compatible with $\zeta_q \neq 0$ but not with $\zeta_q = 0$ and $\gamma_u \neq 0$.

The parameters w_q and v_q can be normalised, since by making the separability preserving transformations

$$x^{n-1} \to ax^{n-1} + b, \tag{5.5}$$

we can take one v_q to be zero and one w_q to be 1. Further normalisations are possible when $E = 0$ because of the extra conformal symmetries K_2 and D. For instance, consider the case when one of the L_q's is I in (5.2). The coordinates on this block are

$$y^b = z^b + \gamma_b (x^{m+1})^2/4, \qquad b \in B_q. \tag{5.6}$$

The dilatation D acts on the coordinates as

$$y^b \to cy^b, \quad t \to c^2 t, \quad c \in \mathbb{R}. \tag{5.7}$$

If this action is combined with the *equivalence* transformations

$$z^b \to cz^b, \qquad x^{m+1} \to c^2 x^{m+1}, \tag{5.8}$$

and $c=\gamma_b^{-1/3}$ then γ_b can be normalised to 1 or 0. If L_q is II on a block, then the same normalisation is possible by similar methods. There are more equivalences possible under the conformal symmetries, but the discussion of these will be postponed until the operators have been determined in §6. It is not possible to simultaneously normalise all the γ_b's. There is a different coordinate system for each vlaue of γ_b. Thus in general, there is an infinite number of R-separable coordinate systems for (*) and the numbering of these systems in the Appendix is for tidiness only.

The following procedure emerges for the construction of all R-separable systems for (*).

A. Construct the graphs representing all separable systems on $\mathbb{R}^m$.

B. For each of these construct all possible boxings.

C. From the discussion above determine all possible σ's compatible with the boxings.

We now go through this procedure for $m=1$.

A. There is only one possible separable system on $\mathbb{R}^1$

(5.9) 

corresponding to the choice of coordinate

(5.10) $$z^1=x^1.$$

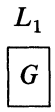
B. There is only one possible boxing:

(5.11)  .

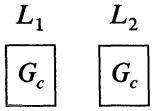
C. The resulting types with their coordinate transformations are listed in Table 1 of the Appendix.

We have displayed possible normalisations of the parameters v_q, w_q in brackets alongside L_q, but have not substituted their values in the coordinate transformations since the unnormalised forms will be needed for the $m=2$ and $m=3$ classifications.

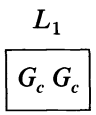
For $m=2$ the separable systems resulting from step A are listed in Table 2 of the Appendix. There are three classes of boxings arising from step B.

(5.12a) 

where G is the elliptic, parabolic or polar system of Table 2 in the Appendix,

(5.12b) 

and

(5.12c) 

where G_c is $\langle\overline{0}\rangle$. The only new systems are those of type (5.12a) as it will be shown that the remaining systems can be derived from the $m=1$ case. The unsplit class a types are

listed in Table 3 of the Appendix. The class b systems are mixtures of the $m=1$ systems

$$\overset{L_1}{\boxed{G_c}} \quad \text{and} \quad \overset{L_2}{\boxed{G_c}} \tag{5.13}$$

where $\sigma_1 \neq \sigma_2$, and these are listed in Table 4 of the Appendix. $\mathbb{R}^2$ has *split* into $\mathbb{R} \oplus \mathbb{R}$ and so these coordinate systems are referred to as *splitting types*. For example, the system 5: II, III ($v_2=0$) in Table 4 has coordinate transformations

$$\begin{aligned} y^1 &= (x^3+v_1)x^1 + \gamma_1/4(x^3+v_1), \\ y^2 &= |x^3|^{1/2}x^2, \\ t &= x^3 \end{aligned} \tag{5.14}$$

that are easily derived from systems II and III in Table 1. The remaining systems of class c are simply combinations of the $m=1$ systems

$$\overset{L_1}{\boxed{G_c}} \quad \text{and} \quad \overset{L_1}{\boxed{G_c}} \tag{5.15}$$

where $\sigma_1=\sigma_2$, that is class b with $L_1=L_2$. They are listed in Table 4 of the Appendix. For example the coordinate transformations for IV$+(v_1=0, w_1^2=1)$ can be derived from those of the system

$$\overset{IV+(v_1=0,\, w_1^2=1)}{\boxed{\langle 0 \rangle}} \tag{5.16}$$

given in Table 1. They are

$$\begin{aligned} y^1 &= \left|(x^3)^2+1\right|^{1/2}x^1, \\ y^2 &= \left|(x^3)^2+1\right|^{1/2}x^2, \\ t &= x^3. \end{aligned} \tag{5.17}$$

Miller (1977) has also classified the R-separable systems for $m=2$ and developed many of their properties but misses splitting types such as (5.14) and (5.17). In general, his classification only includes those mixing types of class b for which one of the coordinates is $y^1=x^1$. These omissions are rectified in Kalnins and Miller (1979).

For general m the R-separable systems have form

$$\overset{L_1}{\boxed{G}}, \tag{5.18}$$

where G is of parabolic or elliptic-type, or they will be a mixture of systems like (5.2). In this case (*) is equivalent to the equations

$$(\Delta_q + 2\varepsilon\partial_t)\Psi_q = T_q\Psi_q, \qquad \Psi = \prod_{q\in Q}\Psi_q, \tag{5.19}$$

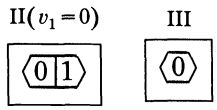
where

$$\Psi_q = \Psi_q(x^b), \qquad b \in B_q. \tag{5.20}$$

All such mixtures can be classified from the lower dimensional equations given in (5.19).

The classification for $m=3$ for the unsplit types given in Table 5 of the Appendix has not appeared before. These systems can of course be derived from our general procedure but we list them for easy access. The splitting types for $m=3$ can be derived from the tables for $m=1$ and $m=2$ given in the Appendix. For example consider

(5.21)



for which the coordinates are

$$\begin{aligned} y^1 &= cx^4 \cosh(x^1)\cos(x^2), \\ y^2 &= cx^4 \sinh(x^1)\sin(x^2), \\ y^3 &= |x^4 + v_3|^{1/2} x^3, \\ t &= x^4, \end{aligned} \tag{5.22}$$

and these can be obtained from the tables for $m=1$ and $m=2$.

6. Operators. Each R-separable system for $(*)$ is characterised as a commuting set of second order differential operators that are in the enveloping algebra of $(*)$. These operators are derived from those representing separable systems on $\mathbb{R}^m$. In this way we will establish the connection between separation of variables for $(*)$ and the symmetry group of $(*)$.

We will first derive the operators for (2.1) from the Killing tensors λ_i for the Hamilton–Jacobi equation (4.1) associated with (2.1). The Killing tensors characterising separable systems for (4.1) are of second order, and the work of Thomas (1946) implies that they are in the enveloping algebra of $e(n,1)$. In other words if the Killing vectors μ_j are a basis for $e(n,1)$ then there are real constants a^{jk} such that

$$\lambda_i = \sum_{j,k} a^{jk} \{\mu_j, \mu_k\}, \qquad 1 \leqq i,j,k \leqq n. \tag{6.1}$$

Recall that $\{\cdot,\cdot\}$ is the symmetric bracket which is defined by

$$\{\mu_j, \mu_k\} = \frac{\mu_j \mu_k + \mu_k \mu_j}{2}. \tag{6.2}$$

By a simple generalisation of the work of Kalnins and Miller (1977) the corresponding operators $\tilde{\lambda}_i$ for (2.1) can be obtained by the identification $\mu_j \to \tilde{\mu}_j$ defined in (2.25) so that

$$\tilde{\lambda}_i = \sum_{j,k} a^{jk} \{\tilde{\mu}_j, \tilde{\mu}_k\}. \tag{6.3}$$

These will also be operators for $(*)$ with the identification $\hat{\varepsilon} \to \varepsilon$ since

$$\tilde{\lambda}_i = \overline{\Psi} = l_i \overline{\Psi} \quad \text{iff} \quad \tilde{\lambda}_i \Psi = l_i \Psi \tag{6.4}$$

where $\overline{\Psi}$ and Ψ are defined in (2.7).

The work of Shapovalov (1979), and Kalnins and Miller (1981) provides the machinery to determine the Killing tensors. They are

$$\lambda_i = \sum_{a=1}^{n-2} \frac{\Phi^{ai}}{\Phi} p_a^2 + 2\frac{\Phi^{(n-1)i}}{\Phi} p_{n-1}p_n + \sum_{a=1}^{n-2} \frac{\Phi^{ai}}{\Phi} A_a^{nn} p_n^2, \qquad i=1,2,\cdots,n-1, \tag{6.5}$$

where (Φ_{ij}) is the Stäckel matrix defined in (4.6). We first find this matrix in terms of the Stäckel matrices for the embedded separable Euclidean systems $d\bar{s}_q^2$. These are the $n(B_q)\times n(B_q)$ dimensional matrices, $(\Phi_q)_{ij}$, such that

$$\bar{g}^{bb} = \frac{\Phi_q^{b1}}{\Phi_q}, \qquad \Phi_q = \det\left((\Phi_q)_{ij}\right). \tag{6.6}$$

The matrix (Φ_{ij}) can be taken as

$$\begin{pmatrix} 0 & \boxed{\Phi_1} & \cdots & 0 & \cdots & 0 \\ \vdots & & \ddots & & & \\ & & 0 & \boxed{\Phi_q} & \cdots & \\ 0 & & & \vdots & \ddots & \\ 1 & \frac{-1}{\sigma_1} & 0 & \cdots & \frac{-1}{\sigma_q} \quad 0 & \cdots \end{pmatrix} \tag{6.7}$$

since this matrix satisfies

$$g^{bb} = \frac{\Phi^{b1}}{\Phi} = \frac{1}{\sigma_q}\frac{\Phi_q^{b1}}{\Phi_q}, \quad \text{for } b \in B_q \tag{6.8}$$

By substituting for (Φ_{ij}) from (6.7) into (6.5), the Killing tensors are

$$\begin{aligned} &\lambda_v = \sum_{b\in B_q} \frac{\Phi_q^{bv}}{\Phi_q} p_b^2 + U_v p_n^2, \\ &v = 1,2,\cdots,m, \\ &\lambda_{m+1} = E, \end{aligned} \tag{6.9}$$

where

$$U_v = \sum_{b\in B_q} \frac{\Phi_q^{bv}}{\Phi_q} A_b^{nn}. \tag{6.10}$$

In (6.9) the terms $\Sigma_{b\in B_q}(\Phi_q^{bv}/\Phi_q)p_b^2$ can be recognised as having the same form as that for the constants of the motion for the separable system $d\bar{s}_q^2$. This fact will guide us in what follows.

The constant of the motion λ_v may also be written

$$\lambda_v = \mathbf{p}^t \Lambda_v \mathbf{p} \tag{6.11}$$

where

$$\mathbf{p}=\begin{pmatrix} p_1 \\ \vdots \\ p_n \end{pmatrix},$$

and

$$\Lambda_v=\begin{pmatrix} 0 & & & & & & \\ & \ddots & & & & 0 & \cdots \\ & & & & \Lambda_v^E & & \\ & & & \iddots & & \ddots & \\ & & 0 & & & & \\ \cdots & & & & & & U_v \end{pmatrix}. \tag{6.12}$$

Here Λ_v^E is a diagonal matrix whose elements as determined by (6.9) are $(\Lambda_v^E)_{bc}=\delta^{bc}\Phi_q^{bv}/\Phi_q$. It is what will be called the matrix form of the Euclidean constants of the motion arising on $d\bar{s}_q^2$. We express λ_v in terms of the y^i coordinates by using the transformation matrix J:

$$J_{ab}=\frac{\partial y^b}{\partial x^a}. \tag{6.13}$$

In the y^i coordinates

$$\lambda_v=\mathbf{P}^t J^t \Lambda_v J \mathbf{P} \tag{6.14}$$

where

$$\mathbf{P}=\begin{pmatrix} P_1 \\ \vdots \\ P_n \end{pmatrix}, \qquad P_i=\partial_{y^i} W. \tag{6.15}$$

and t denotes matrix transposition. Using (6.13) J is given by

$$\begin{pmatrix} \sigma_1^{1/2}J_1 & & \cdots & & & \mathbf{c}_1 & \mathbf{c}_1 \\ & \ddots & & & 0 & \vdots & \vdots \\ & & \sigma_q^{1/2}J_q & & & \mathbf{c}_q & \mathbf{c}_q \\ & 0 & & \ddots & & \vdots & \vdots \\ B_1 & \cdots & B_q & \cdots & & X & Y \\ 0 & \cdots & 0 & \cdots & & \tfrac{1}{2} & \tfrac{1}{2} \end{pmatrix} \tag{6.16}$$

where from (4.30)

$$(\mathbf{c}_q)_a=-\frac{1}{2}\partial_{x^a}\sum_{b\in B_q} G_{bq} \quad \text{and} \quad (J_q)_{ab}=\partial z^b/\partial x^a. \tag{6.17}$$

As our final result is independent of the quantities B_q, X and Y we need not calculate

them. Finally

$$\lambda_v=\sigma_q\mathbf{P}_q^t\tilde{\Lambda}_v^E\mathbf{P}_q-\mathbf{P}_q^t\varepsilon\left[\sigma_q'\sigma_q^{1/2}\tilde{\Lambda}_v^E\mathbf{z}_q+\sigma_q^{1/2}\left(\int\sigma_q^{-3/2}\right)\tilde{\Lambda}_v^E\boldsymbol{\gamma}_q\right]+F_v\varepsilon^2. \tag{6.18}$$

In this equation

$$F_v=4\mathbf{c}_q^t\tilde{\Lambda}_v^E\mathbf{c}_q+U_v,\qquad \tilde{\Lambda}_v^E=J^t\Lambda_u^EJ \tag{6.19a}$$

and $\boldsymbol{\gamma}_v$, $\mathbf{P}_q$ and $\mathbf{z}_q$ are $n(B_q)\times 1$ vectors:

$$(\boldsymbol{\gamma}_q)_b=\gamma_b,\quad(\mathbf{P}_q)_b=P_b,\quad(\mathbf{z}_q)_b=z^b,\quad b\in B_q. \tag{6.19b}$$

From (6.19a) $\tilde{\Lambda}_v^E$ is just Λ_v^E in the z^b coordinates. Equation (6.18) implies that the constants of the motion for $v\in B_q$ do not depend on the structure of (*) on any of the other blocks. This is natural since (*) is equivalent to the set of lower dimensional equations given in (5.19). The problem of finding the operators for a general system has now reduced to the determination of the constants of the motion on one of the blocks

$$\begin{matrix}L_q\\ \boxed{G_q}\end{matrix}. \tag{6.20}$$

Kalnins and Miller (1982a) give specific formulae for the Euclidean constants of the motion $\tilde{\Lambda}_{\lesssim}^E$. On each irreducible block they find that

$$P_b^2,\ M_{ab}^2 \text{ and } \{M_{qb},P_b\},\qquad a,\ b\in B_q, \tag{6.21}$$

are a basis so that

$$\tilde{\Lambda}_v^E=\sum_b A^bP_b^2+\sum_b B^b\{M_{qb},P_b\}+\sum_{a<b}C^{ab}M_{ab}^2,\qquad A^b,\ B^b,\ C^{ab}\in\mathbb{R}. \tag{6.22}$$

The reader is referred to Kalnins and Miller (1982a) for the determination of the constants in (6.22). Since (6.18) provides us with a relation of form

$$\lambda_v=\lambda_v\left(\tilde{\Lambda}_v^E,\sigma_q\right) \tag{6.23}$$

all that has to be done for each of the σ's for $\mathrm{I}\to\mathrm{IV}\pm$, is to determine the images of the Euclidean constants of the motion in (6.21). In each case we will find an expression λ' in the enveloping algebra such that λ is determined to within a term in ε^2 i.e.

$$\lambda=\lambda'+F'\varepsilon^2. \tag{6.24}$$

(In this discussion prime is not the derivative). Since λ is a constant of the motion and λ' is in the enveloping algebra

$$\{E,\lambda\}_P=0=\{E,\lambda'+F'\varepsilon^2\}_P=\{E,F'\varepsilon\}. \tag{6.25}$$

Solving this relation with $E=P_1^2+\cdots+P_{n-1}^2-P_n^2$ we find that F' is a constant, but since ε is already on the orbit

$$\lambda=\lambda' \tag{6.26}$$

in (6.24).

The images of the Euclidean operators are displayed in Table 6.1; however one example will be done in detail to show how they are derived. Consider the term $\{M_{qb}, P_b\}$ which appears in constants of the motion in parabolic-type coordinates. The matrix form $\tilde{\Lambda}^E$ for this term is

$$\begin{pmatrix} & & -z^b/2 & & \\ & & 0 & & \\ & & \cdot & & \\ & & \cdot & & \\ -z^b/2 & 0 \quad \cdots & z^q & \cdots & 0 \\ & & \cdot & & \\ & & \cdot & & \\ & & 0 & & \end{pmatrix} \tag{6.27}$$

as it is easily checked that $\mathbf{P}^t\tilde{\Lambda}^E\mathbf{P} = \{M_{qb}, P_b\}$. From (4.30) and (4.50)

$$\begin{aligned} y^q &= z^q + \gamma_q (x^{n-1})^2/4, \\ y^b &= z^b, \qquad b \neq q. \end{aligned}$$

The calculation for the constant of the motion corresponding to $\{M_{qb}, P_b\}$ goes as follows:

$$\begin{aligned} \lambda\big(\{M_{qb}, P_b\}, \sigma = 1\big) &= \{z^q P_{y^b} - z^b P_{y^q}, P_b\} - \mathbf{P}_q^t x^{n-1} \tilde{\Lambda}^E \gamma_q \varepsilon + F\varepsilon^2 \\ &= \{M_{qb}, P_b\} - \frac{1}{4}\gamma_q (x^{n-1})^2 P_b^2 + \frac{1}{2}\gamma_q x^{n-1} P_b \varepsilon + F'\varepsilon^2 \\ &= \{M_{qb}, P_b\} - \gamma_q B_b^2/4 \end{aligned} \tag{6.28}$$

since $F' = 0$ from the discussion above. The corresponding operator for (2.1) and (*) is obtained via (6.3) but it may be unambiguously written as (6.28) by removing the tildes. Occasionally P_b has been used to represent both the Euclidean constant of the motion $\partial W/\partial z^b$ and $\partial W/\partial y^b$, a constant of the motion for the Hamilton–Jacobi equation (4.1). Sometimes, as in (6.28), we have written P_{y^b} to emphasise the difference.

TABLE 6.1
Images of Euclidean operators.

$\tilde{\Lambda}^E$	L_q			
	I	II	III	IV±
P_b^2	$P_b^2 + \gamma_b \varepsilon B_b$	$(v_q P_b - B_b)^2 + \gamma_b \varepsilon P_b$	$v_q P_b^2 - \{P_b, B_b\}$	$(v_q P_b - B_b)^2 \pm w_q^2 P_b^2$
M_{ab}^2	M_{ab}^2	M_{ab}^2	M_{ab}^2	M_{ab}^2
$\{M_{qb}, P_b\}$	$\{M_{qb}, P_b\} - \gamma_q B_b^2/4$	$\{M_{qb}, v_q P_b - B_b\} - \gamma_q P_b^2/4$	does not occur	does not occur

The results in Table 6.1 can be used to obtain the operators for any m but we will go through these results in detail for $m = 1$ and $m = 2$.

When $m = 1$ the Euclidean operator is P_1^2, and its images can be read from the row containing P_b^2 in Table 6.1 obtaining the results listed in Table 1 of the Appendix.

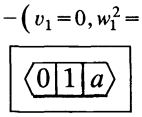
When $m=2$ the operators resulting from the unmixed systems are listed in Table 3 of the Appendix. The operators for the mixing systems are simply those of the component one-dimensional systems. For example the two Euclidean operators for (5.14) are P_1^2 and P_2^2 and by using Table 6.1 their images are

$$(v_1P_1-B_1)^2+\gamma_1\varepsilon P_1, \text{ and } -\{P_1,B_1\}. \tag{6.29}$$

Alternatively the same result is easily obtained from Table 1 of the Appendix.

In the three-dimensional case the operators have not been listed in the Appendix, but the following example will show how they are obtained. Consider system 4 of Table 6 in the Appendix.

(6.30)



The operators for ellipsoidal coordinates are easily derived from the work of Kalnins and Miller (1982a) but we also record them in Table 5. They are

$$\begin{aligned} &P_1^2+P_2^2+P_3^2,\\ &\mathbf{J}\cdot\mathbf{J}+c^2\left[(1+a)P_1^2+aP_2^2+P_3^2\right],\\ &J_2^2+aJ_3^2+c^2aP_1^2. \end{aligned} \tag{6.31}$$

Employing Table 6.1,

$$J_b^2\to J_b^2, \qquad P_b^2\to B_b^2-P_b^2, \qquad b=1,2,3 \tag{6.32}$$

so that the operators for (6.30) are

$$\begin{aligned} &\sum_b\left(B_b^2-P_b^2\right), \mathbf{J}\cdot\mathbf{J}+c^2\left[(1+a)\left(B_1^2-P_1^2\right)+a\left(B_2^2-P_2^2\right)+B_3^2-P_3^2\right],\\ &J_2^2+aJ_3^2+c^2a\left(B_1^2-P_1^2\right). \end{aligned} \tag{6.33}$$

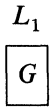
The operators for the $m=3$ mixing types are also computed in the same fashion. For instance the operators for the mixing type (5.21) can be derived from the tables for the one- and two-dimensional cases. They are

$$B_1^2+B_2^2, \quad M_{12}^2+c^2B_1^2 \quad \text{and} \quad v_3P_3^2-\{P_3,B_3\}. \tag{6.34}$$

All R-separable systems for the $m=3$ for (*) have been characterised as commuting sets of second order partial differential operators which are members of the enveloping algebra. An immediate application for this characterisation is now given.

All separable systems on $\mathbb{R}^m$ possess the Casimir operator $\Sigma_{u=1}^m P_u^2$. If $E=0$ and our systems have form

(6.35)



then the operators corresponding to this invariant for each of the σ types are:

$$
\begin{aligned}
&\text{I} && \sum_1^m P_b^2+\gamma_1\varepsilon B_1,\\
&\text{II} && \sum_1^m B_b^2+\gamma_1\varepsilon P_1,\\
&\text{III} && \sum_1^m \{P_b,B_b\},\\
&\text{IV}\pm && \sum_1^m B_b^2\pm\sum_1^m P_b^2.
\end{aligned}
\tag{6.36}
$$

By examining the forms of the operators in (2.25)→(2.26) the following substitutions hold on the solution space of (*) when $E=0$.

$$
\sum_1^m P_b^2\to -2\varepsilon K_{-2},\quad \sum_1^m \{P_b,B_b\}\to\varepsilon D \quad\text{and}\quad \sum_1^m B_b^2\to 2\varepsilon K_2. \tag{6.37}
$$

With these substitutions, the new expressions for the operators in (6.36) are summarised in Table 6.2.

TABLE 6.2

L_1	Operator
I	$\varepsilon[\gamma_1 B_1-2K_{-2}]$
II	$\varepsilon[2K_2+\gamma_1 P_1]$
III	εD
IV±	$\varepsilon[K_2\mp K_{-2}]$

Each of the operators in Table 6.2 is first order and can be diagonalised to reduce (*) by a dimension. This feature is discussed by Miller (1977) for the one and two-dimensional heat equations. His claim, however, that this is a feature of all systems of those dimensions does not hold. For example the operators (6.29) of the splitting type (5.14) can never be made first order.

The knowledge of the operators will now be exploited to find the extra equivalences for $E=0$ that were mentioned in §5. For a system like (6.35) it is possible, as was explained in §5, to normalise γ_1 to 0 or 1. Thus there are the possibilities for L_1:

$$
\begin{aligned}
&1 \quad \text{a.} \quad \text{I} \quad (\gamma_1=0)\\
&\quad\ \ \text{b.} \quad \text{I} \quad (\gamma_1=1)\\
&2 \quad \text{a.} \quad \text{II} \quad (v_1=0,\gamma_1=0)\\
&\quad\ \ \text{b.} \quad \text{II} \quad (v_1=0,\gamma_1=1)\\
&3 \quad\ \ \text{III} \quad (v_1=0)\\
&4 \quad \text{a.} \quad \text{IV}- \quad (v_1=0,w_1^2=-1)\\
&\quad\ \ \text{b.} \quad \text{IV}+ \quad (v_1=0,w_1^2=1).
\end{aligned}
\tag{6.38}
$$

The extra equivalences occur under the action of the operator $A_3=K_{-2}-K_2$. The

adjoint action of this operator is given in Table 6.3.

TABLE 6.3
Adjoint action of A_3.

λ	$e^{s\,\mathrm{Ad}\,A_3}\lambda$
P_a	$\cos(s)P_a+\sin(s)B_a$
M_{ab}	M_{ab}
B_a	$\cos(s)B_a-\sin(s)P_a$
A_1	$A_1\cos(2s)-A_2\sin(2s)$
A_2	$A_2\cos(2s)-A_1\sin(2s)$

The adjoint action of A_3 on any element $\{L_1,L_2\}$ of the enveloping algebra is easily shown to be

$$e^{s\,\mathrm{Ad}\,A_3}(\{L_1,L_2\})=\{e^{s\,\mathrm{Ad}\,A_3}(L_1),e^{s\,\mathrm{Ad}\,A_3}(L_2)\} \tag{6.39}$$

by using (2.29). Consider the action of A_3 when $s=-\pi/4$ and for convenience let $J=e^{-\pi A_3/4}$. Then from (2.29)

$$e^{-\pi\,\mathrm{Ad}\,A_3/4}(L)=JLJ^{-1}. \tag{6.40}$$

The Euclidean operators for (∗) are

$$\tilde{\lambda}_v^E=\sum_b A_v^b P_b^2+\sum_b B_v^b\{M_{qb},P_b\}+\sum_{a<b} C_v^{ab}M_{ab}^2 \tag{6.41}$$

where $v=1,\cdots,m+1$. For case 3 in (6.38), $B_v^b=0$ always, since such terms can only appear for parabolic type coordinates. The corresponding operators as obtained from Table 6.1 are

$$\lambda_v=-\sum_b A_v^b\{P_b,B_b\}+\sum_{a<b} C_v^{ab}M_{ab}^2. \tag{6.42}$$

If the action of J is applied to these operators using Table 6.3 we obtain

$$\lambda'_v=-\sum_b A_v^b\left\{\frac{P_b}{\sqrt{2}}-\frac{B_b}{\sqrt{2}},\frac{B_b}{\sqrt{2}}+\frac{P_b}{\sqrt{2}}\right\}+\sum_{a<b} C_v^{ab}M_{ab}^2 \tag{6.43}$$

$$=\sum_b A_v^b(B_b^2-P_b^2)/2+\sum_{a<b} C_v^{ab}M_{ab}^2$$

since $P_b\to P_b/\sqrt{2}-B_b/\sqrt{2}$, $B_b\to B_b/\sqrt{2}+P_b/\sqrt{2}$ and $M_{ab}\to M_{ab}$ under the action of J. The expression obtained in (6.43) is precisely the one that would have been obtained from the same Euclidean operators for system 4a in (6.38). In similar fashion it can be shown that the systems 1a and 1b are equivalent to the systems 2a and 2b respectively under the action of

$$J^2=e^{-\pi A_3/2}. \tag{6.44}$$

We have demonstrated that the systems in (6.38) collapse to the systems 1a, 1b, 3 and 4b. The action of J on solutions of (*) is

$$
\begin{aligned}
&J\Phi(\mathbf{y},t)=\left[\frac{\sqrt{2}}{(1+t)}\right]^{m/2}\exp\left[\frac{\varepsilon\mathbf{y}\cdot\mathbf{y}}{2(1+t)}\right]\Phi\left(\frac{\sqrt{2}\,\mathbf{y}}{t+1},\frac{t-1}{t+1}\right),\\
&J^2\Phi(\mathbf{y},t)=t^{-m/2}\exp\left[\frac{\varepsilon\mathbf{y}\cdot\mathbf{y}}{2t}\right]\Phi\left(\frac{\mathbf{y}}{t},\frac{-1}{t}\right),\\
&J^4\Phi(\mathbf{y},t)=-\Phi(-\mathbf{y},t),\\
&J^8\Phi(\mathbf{y},t)=\Phi(\mathbf{y},t).
\end{aligned}
\tag{6.45}
$$

The expression for J^2 is particularly notable: it is the *Appell* transform, and its importance for the theory of the Heat equation is discussed in Widder (1975). The above work is a generalisation of that of Miller (1977) for the one and two-dimensional heat equations.

In conclusion, it is possible to exploit the conformal symmetries for unsplit systems when $E=0$ to show that some systems that look different are actually equivalent under the action of the Schrödinger group.

7. *R*-separable solutions. We now investigate the R-separable solutions of (*). To accomplish this, it is necessary to determine both the R-separation factors and the *separation equations*. The separation equations for (*) are the ordinary differential equations determining the Ψ_j functions in (2.5). From (2.7) these are the same as the corresponding functions for the Helmholtz equation (2.1). Therefore this equation is used to find the separation equations for (*). The form of (2.1) in the separable coordinates x^i can be found by using this equations local coordinate description given in (2.14). It is

$$
\begin{aligned}
\sum_{q\in Q}\frac{1}{\sigma_q}\sum_{b\in B_q}&\left[\bar{g}_q^{-1/2}\partial_b\left(\bar{g}^{bb}\bar{g}_q^{1/2}\partial_b\Psi\right)+\bar{g}^{bb}A_b^{nn}\partial_{nn}\Psi\right]\\
&+2\partial_{(n-1)n}\Psi+\frac{1}{2}\sum_{q\in Q}n(B_q)(\log\sigma_q)'\partial_n\Psi=E\Psi
\end{aligned}
\tag{7.1}
$$

where

$$
\bar{g}_q=\det(\bar{g}_{bc}),\qquad b,c\in B_q,
\tag{7.2}
$$

and A_b^{nn} is defined in (4.14). When $\Psi=\Pi_{i=1}^n\Psi_i(x^i)$ is substituted into (7.1) and the resulting equation is divided by Ψ we obtain the separation equations:

$$
\Psi_n'=\varepsilon\Psi_n,
\tag{7.3}
$$

and

$$
2\varepsilon\Psi_{n-1}'+\sum_{q\in Q}\frac{s_q}{\sigma_q}\Psi_{n-1}+\frac{\varepsilon}{2}\sum_{q\in Q}n(B_q)(\log\sigma_q)'\Psi_{n-1}=E\Psi_{n-1}.
\tag{7.4}
$$

Here s_q is the eigenvalue of the operator S_q whose action on Ψ is

$$
S_q\Psi=\sum_{b\in B_q}\left[\bar{g}_q^{-1/2}\partial_b\left(\bar{g}^{bb}\bar{g}_q^{1/2}\partial_b\Psi\right)+(\varepsilon)^2\bar{g}^{bb}A_b^{nn}\Psi\right]=s_q\Psi.
\tag{7.5}
$$

The remaining separation equations are derived from (7.5) by noticing it is the Helmholtz equation on $d\bar{s}_q^2$ with an extra term

$$\sum_{b\in B_q} (\varepsilon)^2 \bar{g}^{bb} A_b^{nn} \Psi. \tag{7.6}$$

By a simple modification of Kalnins and Miller's results for $\mathbb{R}^{n(B_q)}$ we obtain these equations.

The one-dimensional elliptic case is an exception to the general construction of the separation equations that we are about to give. The separation equation for this case is

$$\frac{d^2\Psi_1}{(dx^1)^2} + \left\{ \varepsilon^2 \left[\zeta_1 (x^1)^2/4 + \gamma_1 x^1 \right] - s_1 \right\} \Psi_1 = 0. \tag{7.7}$$

Returning to our general treatment, we can confine ourselves to a block E_r of form (4.32A) or (4.32B). For simplicity on each of these irreducible blocks the coordinates x^r, $x^{r+1}, \cdots, x^{r+n_r-1}$ are relabelled as x^1, $x^2, \cdots, x^{n_r}$ just as in our discussion of the coordinate transformations at the end of §4. The additional term contributed to the separation equations by (7.6) for the elliptic-type A coordinates x^a is

$$Z_a = -\frac{(\varepsilon)^2 c_r^4 \zeta_q}{16} (x^a)^{N_r-1} \left(x^a - \sum_{b=1}^{N_r} e_b \right), \qquad a = 1, 2, \cdots, N_r. \tag{7.8A}$$

For the parabolic-type B coordinates x^a this term is

$$Z_a = \frac{(\varepsilon)^2 c_r^3 \gamma_1}{8} (x^a)^{N_r-1} \left(x^a + \sum_{b=1}^{N_r-1} e_b \right), \qquad a = 1, 2, \cdots, N_r. \tag{7.8B}$$

The separation equations in both of these cases are

$$(P_a/Q_a)^{1/2} \frac{d}{dx^a} \left[(P_a Q_a)^{1/2} \frac{d}{dx^a} \Psi_a \right] + \left\{ \sum_{b=1}^{N_r} \frac{\Pi_{c\neq b}(e_b - e_c)}{(x^a - e_b)} t_b + \sum_{b=1}^{N_r} l_b (x^a)^{N_r-b} + Z_a \right\} \Psi_a = 0. \tag{7.9}$$

In this last equation

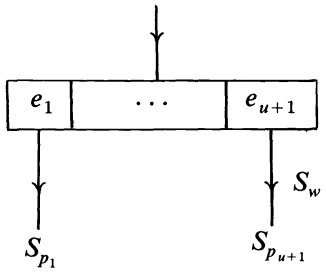
$$P_a = \prod_{b=1}^{N_r} (x^a - e_b), \qquad Q_a = \prod_{b=1}^{N_r} (x^a - e_b)^{p_b}$$
$$t_b = \begin{cases} 0, & p_b = 0 \\ j_b(j_b + p_b - 1), & p_b \neq 0 \end{cases} \tag{7.10}$$

where p_b is the dimension of the sphere S_{p_b} attached to e_b (see (4.32A) and (4.32B)). If there is no sphere attached, then $p_b = 0$. From (7.5) there is no loss in assuming $l_1 = -s_1$. The constant t_b is the eigenvalue of the Helmholtz equation on the sphere S_{p_b}.

The form of t_b given in (7.10) appears in Kalnins and Miller (1982a) and is derived from the spectral theory of the Sphere.

We now consider the separation equations for the remaining variables i.e. those on the attached spheres S_{p_b}. As was mentioned in §4 the separable systems for such spaces have been classified by Kalnins and Miller (1982a). If we trace down one of the tree graphs they use to represent these systems, it will have form (see (4.38))

(7.11)



The separation equations for the variables $x^1, \cdots, x^u$ on S_w are

$$(7.12)\qquad \left(\frac{P_a}{Q_a}\right)^{1/2}\frac{d}{dx^a}\left[(P_aQ_a)^{1/2}\frac{d}{dx^a}\Psi_a\right] + \left\{\sum_{b=1}^{u+1}\frac{\Pi_{c\neq b}(e_b-e_c)}{(x^a-e_b)}t_b+\sum_{b=1}^{u}l_b(x^a)^{u-b}\right\}\Psi_a=0, \qquad a=1,2,\cdots,u,$$

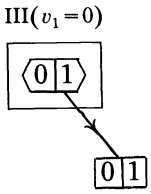
with the same definitions for P_a, Q_a and t_b except that the number of elliptic or parabolic coordinates N_r is replaced by u. We can of course take $-l_1$ as the value of the Helmholtz equation on the sphere S_w and via the spectral theory to be $-\mu(\mu+w-1)$ (see Talman (1968)). These separation equations are unaffected by the terms V_q and σ_q, and are just the same as they would be on Euclidean space.

The separation equation for x^{n-1}, (7.4), can be directly integrated:

$$(7.13)\qquad \Psi_{n-1}=\left\{\prod_{q\in Q}\sigma_q^{-n(B_q)/4}\right\}\exp\left(\frac{1}{2\varepsilon}\left[Ex^{n-1}-\sum_{q\in Q}s_q\int\frac{dx^{n-1}}{\sigma_q}\right]\right).$$

As an example consider system 9 in Table 6 of the Appendix:

(7.14)



The separation equations for this system are

$$(7.15)\quad (x^a)^{1/2}\frac{d}{dx^a}\left[x^a(x^a-1)\frac{d}{dx^a}\Psi_a\right] + \left\{\frac{j_2^2}{(x^a-1)}+-s_1x^a+l_2+\frac{(\varepsilon)^2c_r^4}{16}x^a(x^a-1)\right\}\Psi_a=0, \qquad a=1,2,$$

and

$$[x^3(x^3-1)]^{1/2}\frac{d}{dx^3}\left([x^3(x^3-1)]^{1/2}\frac{d}{dx^3}\Psi_3\right)-j_2^2\Psi_3=0. \tag{7.16}$$

Here s_1, l_2 and j_2^2 are the separation constants. These coordinates are in standard form but they could easily be transformed to the more familiar expressions in terms of cos, sinh etc. that appear in the Appendix. Indeed letting $x^3=\sin^2\vartheta$, (7.16) becomes

$$\frac{1}{4}\frac{d^2\Psi_\vartheta}{d\vartheta^2}+j_2^2\Psi_\vartheta=0.$$

We also have from (7.13)

$$\Psi_4=|x^4|^{-3/4-s_1/2\varepsilon}e^{Ex^4/2\varepsilon} \tag{7.17}$$

In order to fully determine the R-separable solutions, we now find the R-separation factor R. From (2.9) and (4.30)

$$R=-\varepsilon f=\varepsilon\sum_{q\in Q}\sum_{b\in B_q}G_{bq}. \tag{7.18}$$

If we define

$$R_r=\varepsilon\sum_{b\in E_r}G_{bq}, \tag{7.19}$$

then

$$R=\sum_{q\in Q}\sum_{r\in\mathscr{R}}R_r, \tag{7.20}$$

as a result the essential structure of the R-separation factor is solely dependent on the structure of each of the irreducible blocks E_r. Employing (4.50), it can be assume that $\gamma_{r+1}=\gamma_{r+2}=\cdots=0$, with γ_r being nonzero and zero in the parabolic and elliptic cases respectively. By using (7.18) we obtain

$$R_r=\varepsilon\left\{\frac{\sigma_q'}{4}\sum_{b\in E_r}(z^b)^2+\frac{1}{2}\gamma_r z^r\sigma_q^{1/2}\int\sigma_q^{-3/2}\right\}+\frac{\varepsilon\gamma_r^2\delta_{rq}}{8}\int\left(\int\sigma_q^{-3/2}\right)^2. \tag{7.21}$$

The last term of R_r is a function of x^{n-1} alone and so only contributes to trivial R-separation and can be absorbed in $\Psi_{n-1}(x^{n-1})$. An expression for $\Sigma(z^b)^2$ term in R_r in the separable coordinates x^k can be easily obtained from (4.41), (4.43A) and (4.43B). The expression for z^r in the x^k can be found from (4.50) since only parabolic coordinates can correspond to the linear potential term $\gamma_r z^r$. Thus it is always possible to give the R-factors in the x^k coordinates. These general expressions are summarised in Reid (1984). In each case the R-factor is independent of the variables on the spheres S_{p_b}. The R-separation factors corresponding to the unmixed types $m=1,2,3$ are given in the Appendix. The R-separation factors for the two-dimensional mixing types may be obtained simply from those for the one dimension. Similar comments apply for $m=3$.

In one and two dimensions separation equations can be solved to give known special functions. Miller (1977) lists these results and investigates some of their properties (e.g. bases and overlaps for the Schrödinger equation). In higher dimensions, however, the separation equations lead to new special functions about which little is known. Development of the properties of these functions is a problem for future research.

Conclusion. All R-separable coordinate systems have been classified and characterised in terms of the symmetry group of equation (∗). In a subsequent article we will investigate the applications of this group theoretical characterisation. Moving boundary value problems will be considered. It will be shown how the results for the potential free Schrödinger equation can be applied to equations in which the potential term need not vanish. Boyer (1974), (1976) has classified all time-independent potentials for the Schrödinger equation which admit symmetry groups. Exact forms were obtained for those potentials invariant under the Schrödinger algebra of maximal dimension. In particular he has shown how these cases: the potential free Schrödinger equation, the linear potential Schrödinger equation and the harmonic oscillator Schrödinger equations are all equivalent under the action of the Schrödinger group. Miller (1977) has shown how this equivalence is connected with the separability properties of the one and two dimensional Schrödinger equations: we will generalise his result. The programme initiated by Kalnins and Miller for the one and two-dimensional Schrödinger equations will also be pursued: deriving results about the functions arising as separated solutions of (∗), using the group theoretical characterisations of these systems.

Appendix. The main purpose of this appendix is to summarise in tabular form results concerning R-separation for $m=1,2$ and 3.

In all the tables $\sigma_{\mathrm{II}} \to \sigma_{\mathrm{IV}\pm}$ are the functions of Table 4.1, that is

$$\sigma_{\mathrm{II}} = (x^{m+1}+v)^2,$$
$$\sigma_{\mathrm{III}} = |x^{m+1}+v|,$$

and

$$\sigma_{\mathrm{IV}\pm} = \left|(x^{m+1}+v)^2 \pm w^2\right|, \qquad x^{m+1}=t. \tag{A1}$$

As we have already mentioned

$$\tau = \operatorname{sign}\left((x^{m+1}+v)^2 \pm w^2\right). \tag{A2}$$

In the tables for the unsplit systems we have ambiguously removed the subscripts from the parameters so that $v_1 \to v$, $w_1 \to w$ and $\gamma_1 \to \gamma$.

The general forms for elliptic and parabolic coordinates in (4.35) can be transformed to give their more familiar appearance in low dimensions. For example we make the transformations

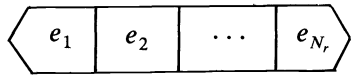
$$\sqrt{x^1} \to \cosh(x^1), \qquad \sqrt{x^2} \to \cos(x^2) \tag{A3}$$

to give elliptic coordinates their "usual" appearance in Table 2. It must be remembered, however, that in higher dimensions this is not always possible. For cases like the ellipsoidal coordinates of Table 5 the general forms of (4.35) must be used. In Table 5, $J_1 = M_{32}$, $J_2 = M_{13}$, $J_3 = M_{21}$ and $\mathbf{J}\cdot\mathbf{J} = \sum_{i=1}^{3} J_i^2$.

In the tables some of the parameters e_r have been normalised. In general by making the transformations

(A4) $$x^i \to ax^i + b, \qquad e_i \to ae_i + b$$

an elliptic block

(A5) 

and become

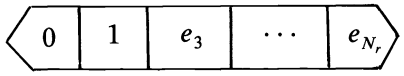
(A6) 

TABLE 1

R-separable coordinates and operators for $\left(\partial_{y^1}^2 + 2\varepsilon\partial_t\right)\Psi = E\Psi$.

	Graph	Coordinates $\{y^1, t\}$ $t = x^2$ for all systems	Operator R-separation factor R
1.	I 	$y^1 = x^1 + \gamma(x^2)^2/4$	$P_1^2 + \gamma\varepsilon B_1$ $R = \varepsilon\gamma x^1 x^2/2$
2.	II$(v=0)$ 	$y^1 = \sigma_{\mathrm{II}}^{1/2} x^1 + \gamma/4(x^2+v)$	$(vP_1 - B_1)^2 + \gamma\varepsilon P_1$ $R = \frac{1}{2}\varepsilon\left[(x^2+v)(x^1)^2 - \gamma x^1/4(x^2+v)\right]$
3.	III$(v=0)$ 	$y^1 = \sigma_{\mathrm{III}}^{1/2} x^1$	$vP_1^2 - \{P_1, B_1\}$ $R = 0$
4.	IV$\pm(v=0, w^2=1)$ 	$y_\pm^1 = \sigma_{\mathrm{IV}\pm}^{1/2} x^1$	$(vP_1 - B_1)^2 \pm w^2 P_1^2$ $R = \tau_\pm \varepsilon (x^1)^2 x^2/2$

TABLE 2

Separable coordinates and operators for $p_1^2 + p_2^2 = E$.

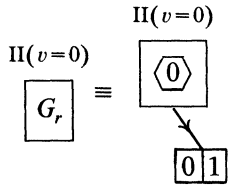
Name & graph	Coordinates	Operator
Elliptic 	$z_e^1 = c\cosh(x^1)\cos(x^2)$ $z_e^2 = c\sinh(x^1)\sin(x^2)$	$M_{12}^2 + c^2 P_1^2$
Parabolic 	$z_p^1 = \frac{1}{2}\left[(x^1)^2 - (x^2)^2\right]$ $z_p^2 = x^1 x^2$	$\{M_{12}, P_2\}$
Polar 	$z_r^1 = x^1\cos(x^2)$ $z_r^2 = x^1\sin(x^2)$	M_{12}^2
Cartesian 	$z^1 = x^1$ $z^2 = x^2$	P_1^2

Analogous remarks apply to parabolic blocks and to those representing separable systems on the spheres S_{p_b}. As a final remark, the split types for $m=3$ may be obtained from the results for $m=1$ and $m=2$ (i.e. from Tables 1 and 3). For a brief discussion see (5.19)→(5.22).

Acknowledgments. This article is the culmination of the early work of Charles Boyer, Ernest Kalnins and Willard Miller. To all of these people, my thanks. I would like to thank Ernest Kalnins especially, for his time, suggestions and encouragement.

Notes.

Table 3. The results of Table 2 have been used to simplify the presentation of this table. For example in system 8



and $y^u=\sigma_{\mathrm{II}}^{1/2}z_r^u$ is

$$y^1=|x^3+v|x^1\cos(x^2),$$
$$y^2=|x^3+v|x^1\sin(x^2).$$

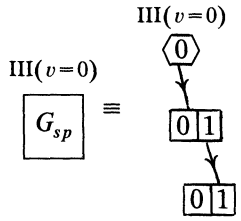
(Recall that $t=x^3$ for these systems.)

Table 4. An entry L_1 L_2 in this table corresponds to the split coordinate system

$$y^1=\sigma_{L_1}^{1/2}x^1+\frac{1}{2}\gamma_1\iint\sigma^{-3/2},$$
$$y^2=\sigma_{L_2}^{1/2}x^2+\frac{1}{2}\gamma_2\iint\sigma^{-3/2},$$
$$t=x^3,\qquad L_i\colon \mathrm{I}\to\mathrm{IV}\pm.$$

An entry L_1 in part c of Table 4 has coordinate transformations as above but with $L_1=L_2$.

Table 6. In analogy to the case $m=2$ we have used the results of Table 5 to simplify the presentation of Table 6. For instance in system 19



and $y^u=\sigma_{\mathrm{III}}^{1/2}z_{sp}^u$ is

$$y^1=|x^4+v|^{1/2}x^1\cos(x^2),$$
$$y^2=|x^4+v|^{1/2}x^1\sin(x^2)\cos(x^3),$$
$$y^3=|x^4+v|^{1/2}x^1\sin(x^2)\sin(x^3).$$

TABLE 3

Unsplit R-separable coordinates and operators for $(\Delta_2 + 2\varepsilon\partial_t)\Psi = E\Psi$.

	Graph	Coordinates $\{y^u, t\}$ $t = x^3$ for all systems	Operators *R*-separation factor *R*
1.	I 	$y^u = z_e^u$	$P_1^2 + P_2^2$ $M_{12}^2 + c^2 P_1^2$ $R = 0$
2.	II($v=0$) 	$y^u = \sigma_{\text{II}}^{1/2} z_e^u$	$(vP_1 - B_1)^2 + (vP_2 - B_2)^2$ $M_{12}^2 + c^2(vP_1 - B_1)^2$ $R = \varepsilon c^2 x^3(\cosh^2(x^1) + \cos^2(x^2))/2$
3.	III($v=0$) 	$y^u = \sigma_{\text{III}}^{1/2} z_e^u$	$v(P_1^2 + P_2^2) - \{P_1, B_1\} - \{P_2, B_2\}$ $M_{12}^2 + c^2(vP_1^2 - \{P_1, B_1\})$ $R = 0$
4.	IV$\pm$($v=0, w^2=1$) 	$y_\pm^u = \sigma_{\text{IV}}^{1/2} z_e^u$	$(vP_1 - B_1)^2 + (vP_2 - B_2)^2 \pm w^2(P_1^2 + P_2^2)$ $M_{12}^2 + c^2[(vP_1 - B_1)^2 \pm w^2 P_1^2]$ $R = \varepsilon\tau_\pm c^2 x^3(\cosh^2(x^1) + \cos^2(x^2))/2$
5.	I 	$y^1 = z_p^1 + \gamma(x^3)^2/4$ $y^2 = z_p^2$	$P_1^2 + P_2^2 + \gamma\varepsilon B_1$ $\{M_{12}, P_2\} - \gamma B_2^2/4$ $R = \varepsilon\gamma x^3((x^1)^2 - (x^2)^2)/4$
6.	II($v=0$) 	$y^1 = \sigma_{\text{II}}^{1/2} z_p^1 + \gamma/4(x^3 + v)$ $y^2 = \sigma_{\text{II}}^{1/2} z_p^2$	$(vP_1 - B_1)^2 + (vP_2 - B_2)^2 + \gamma\varepsilon P_1$ $\{M_{12}, vP_2 - B_2\} - \gamma P_2^2/4$ $R = \frac{\varepsilon}{8}[(x^3 + v)((x^1)^2 + (x^2)^2)^2 - \frac{\gamma((x^1)^2 - (x^2)^2)}{2(x^3 + v)}]$
7.	I 	$y^u = z_r^u$	$P_1^2 + P_2^2$ M_{12}^2 $R = 0$
8.	II($v=0$) 	$y^u = \sigma_{\text{II}}^{1/2} z_r^u$	$(vP_1 - B_1)^2 \quad (vP_2 - B_2)^2$ M_{12}^2 $R = \varepsilon x^3(x^1)^2/2$
9.	III($v=0$) 	$y^u = \sigma_{\text{III}}^{1/2} z_r^u$	$v(P_1^2 + P_2^2) - \{P_1, B_1\} - \{P_2, B_2\}$ M_{12}^2 $R = 0$
10.	IV$\pm$($v=0, w^2=1$) 	$y_\pm^u = \sigma_{\text{IV}}^{1/2} z_r^u$	$(vP_1 - B_1)^2 + (vP_2 - B_2)^2 \pm w^2(P_1^2 + P_2^2)$ M_{12}^2 $R = \varepsilon\tau_\pm x^3(x^1)^2/2$

TABLE 4

Split R-separable coordinates for $(\Delta_2 + 2\varepsilon\partial_t)\Psi = E\Psi$.

Class b splitting types			
1. 4. 7. 9.	I II($v_2=0$) II($v_1=0$) II($v_2\neq 0$) III($v_1=0$) III($v_2\neq 0$) IV± IV±	2. I III($v_2=0$) 5. II III($v_2=0$) 8. III IV±($v_2=0, w_2^2=1$) 10. IV± IV∓	3. I IV±($v_2=0, w_2^2=1$) 6. II IV±($v_2=0, w_2^2=1$)
Class c splitting types			
1. I	2. II($v_1=0$)	3. III($v_1=0$)	4. IV$_\pm$($v_1=0, w_1^2=1$)

TABLE 5

Unsplit separable coordinates and operators for $\mathbf{p}_1^2 + \mathbf{p}_2^2 + \mathbf{p}_3^2 = \mathbf{E}$.

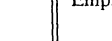
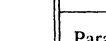
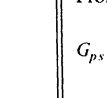
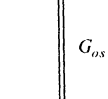
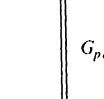
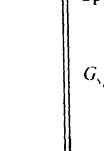
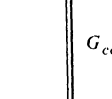
Name & graph	Coordinates	Operators
Ellipsoidal G_{eo} = 	$z^1_{eo} = c[\frac{x^1x^2x^3}{a}]^{1/2}$ $z^2_{eo} = c[\frac{(x^1-1)(x^2-1)(x^3-1)}{(1-a)}]^{1/2}$ $z^3_{eo} = c[\frac{(x^1-a)(x^2-a)(x^3-a)}{a(a-1)}]^{1/2}$ $0<x^1<1<x^2<a<x^3$	$\mathbf{J}\cdot\mathbf{J} + c^2[(1+a)P_1^2 + aP_2^2 + P_3^2]$ $J_2^2 + aJ_3^2 + c^2aP_1^2$
Paraboloidal G_{po} = 	$z^1_{po} = \frac{1}{2}c(x^1+x^2+x^3-1)$ $z^2_{po} = c[-x^1x^2x^3]^{1/2}$ $z^3_{po} = c[(x^1-1)(x^2-1)(x^3-1)]^{1/2}$ $x^1<0<x^2<1<x^3$	$\{J_2,P_3\}-\{J_3,P_2\}+c(P_1^2+P_3^2)$ $J_1^2 + c\{J_3,P_2\}$
Prolate spheroidal G_{ps} = 	$z^1_{ps} = c\cosh(x^1)\cos(x^2)$ $z^2_{ps} = c\sinh(x^1)\sin(x^2)\cos(x^3)$ $z^3_{ps} = c\sinh(x^1)\sin(x^2)\sin(x^3)$	$\mathbf{J}\cdot\mathbf{J} - c^2(P_2^2+P_3^2)$ J_1^2
Oblate spheroidal G_{os} = 	$z^1_{os} = c\cosh(x^1)\cos(x^2)\cos(x^3)$ $z^2_{os} = c\cosh(x^1)\cos(x^2)\sin(x^3)$ $z^3_{os} = c\sinh(x^1)\sin(x^2)$	$\mathbf{J}\cdot\mathbf{J} + c^2(P_2^2+P_3^2)$ J_1^2
Parabolic G_{pa} = 	$z^1_{pa} = \frac{1}{2}[(x^1)^2-(x^2)^2]$ $z^2_{pa} = x^1x^2\cos(x^3)$ $z^2_{pa} = x^1x^2\sin(x^3)$	$\{J_2,P_3\}-\{J_3,P_2\}$ J_1^2
Spherical G_{sp} = 	$z^1_{sp} = x^1\cos(x^2)$ $z^2_{sp} = x^1\sin(x^2)\cos(x^3)$ $z^3_{sp} = x^1\sin(x^2)\sin(x^3)$	$\mathbf{J}\cdot\mathbf{J}$ J_1^2
Conical G_{co} = 	$z^1_{co} = x^1[\frac{x^2x^3}{a}]^{1/2}$ $z^2_{co} = x^1[\frac{(x^2-1)(x^3-1)}{1-a}]^{1/2}$ $z^3_{co} = x^1[\frac{(x^2-a)(x^3-a)}{a(a-1)}]^{1/2}$ $x^1>0, 0<x^2<1<x^3<a$	$\mathbf{J}\cdot\mathbf{J}$ $J_2^2 + aJ_3^2$

TABLE 6

Unsplit R-separable coordinates and R-factors for $(\Delta_3 + 2\varepsilon\partial_t)\Psi = E\Psi$.

	Graph	Coordinates $\{y^u, t\}$ $t = x^4$ for all systems	*R*-separation factor *R*
1.	I 	$y^u = z^u_{eo}$	0
2.	II($v=0$) 	$y^u = \sigma_{\mathrm{II}}^{1/2} z^u_{eo}$	$\frac{1}{2}\varepsilon c^2 x^4 (x^1 + x^2 + x^3)$
3.	III($v=0$) 	$y^u = \sigma_{\mathrm{III}}^{1/2} z^u_{eo}$	0
4.	IV$^+$ ($v=0, w^2=1$) 	$y^u_{\pm} = \sigma_{\mathrm{IV}\pm}^{1/2} z^u_{eo}$	$\frac{1}{2}\varepsilon\tau_{\pm} c^2 x^4 (x^1 + x^2 + x^3)$
5.	I 	$y^1 = z^1_{po} + \gamma (x^4)^2/4$ $y^2 = z^2_{po}$ $y^3 = z^3_{po}$	$\varepsilon c \gamma x^4 (x^1 + x^2 + x^3)/4$
6.	II($v=0$) 	$y^1 = \sigma_{\mathrm{II}}^{1/2} z^1_{po} + \gamma/4 x^4$ $y^2 = \sigma_{\mathrm{II}}^{1/2} z^2_{po}$ $y^3 = \sigma_{\mathrm{II}}^{1/2} z^3_{po}$	$\frac{1}{2}\varepsilon\{c^2 x^4 [2\sum_1^3 (x^i)^2 - (\sum_1^3 x^i)^2 + 6\sum_1^3 x^i] - \gamma c \sum_1^3 x^i/8x^4\}$
7.	I 	$y^u = z^u_{ps}$	0
8.	II($v=0$) 	$y^u = \sigma_{\mathrm{II}}^{1/2} z^u_{ps}$	$\varepsilon c^2 x^4 (\cosh^2(x^1) + \cos^2(x^2))/2$
9.	III($v=0$) 	$y^u = \sigma_{\mathrm{III}}^{1/2} z^u_{ps}$	0
10.	IV$^-$ ($v=0, w^2=1$) 	$y^u_{\pm} = \sigma_{\mathrm{IV}\pm}^{1/2} z^u_{ps}$	$\varepsilon\tau_{\pm} c^2 x^4 (\cosh^2(x^1) + \cos^2(x^2))/2$
11.	I 	$y^u = z^u_{os}$	0
12.	II($v=0$) 	$y^u = \sigma_{\mathrm{II}}^{1/2} z^u_{os}$	$\varepsilon c^2 x^4 (\cosh^2(x^1) + \cos^2(x^2))/2$

Table 6 (continued).

	Graph	Coordinates $\{y^u, t\}$ $t=x^4$ for all systems	R-separation factor R
13.	III($v=0$) G_{os}	$y^u=\sigma_{\mathrm{III}}^{1/2}z^u_{os}$	0
14.	IV$\pm$ ($v=0, w^2=1$) G_{os}	$y^u_{\pm}=\sigma_{\mathrm{IV}\pm}^{1/2}z^u_{os}$	$\varepsilon\tau_{\pm}c^2x^4(\cosh^2(x^1)+\cos^2(x^2))/2$
15.	IV$\pm$ ($v=0, w^2=1$) G_{pa}	$y^1=z^1_{pa}+\gamma(x^4)^2/4$ $y^2=z^2_{pa}$ $y^3=z^3_{pa}$	$\varepsilon\gamma x^4[(x^1)^2-(x^2)^2]/4$
16.	II($v=0$) G_{pa}	$y^1=\sigma_{\mathrm{II}}^{1/2}z^1_{pa}+\gamma/4x^4$ $y^2=\sigma_{\mathrm{II}}^{1/2}z^2_{pa}$ $y^3=\sigma_{\mathrm{II}}^{1/2}z^3_{pa}$	$\frac{\varepsilon}{8}\left[x^4\left((x^1)^2+(x^2)^2\right)^2 - \frac{\gamma\left((x^1)^2-(x^2)^2\right)}{2x^4}\right]$
17.	I G_{sp}	$y^u=z^u_{sp}$	0
18.	II($v=0$) G_{sp}	$y^u=\sigma_{\mathrm{II}}^{1/2}z^u_{sp}$	$\frac{1}{2}\varepsilon x^4(x^1)^2$
19.	III($v=0$) G_{sp}	$y^u=\sigma_{\mathrm{III}}^{1/2}z^u_{sp}$	0
20.	IV$\pm$ ($v=0, w^2=1$) G_{sp}	$y^u_{\pm}=\sigma_{\mathrm{IV}\pm}^{1/2}z^u_{sp}$	$\frac{1}{2}\varepsilon\tau_{\pm}x^4(x^1)^2$
21.	I G_{co}	$y^u=z^u_{co}$	0
22.	II($v=0$) G_{co}	$y^u=\sigma_{\mathrm{II}}^{1/2}z^u_{co}$	$\frac{1}{2}\varepsilon x^4(x^1)^2$
23.	III($v=0$) G_{co}	$y^u=\sigma_{\mathrm{III}}^{1/2}z^u_{co}$	0
24.	IV$\pm$ ($v=0, w^2=1$) G_{co}	$y^u_{\pm}=\sigma_{\mathrm{IV}\pm}^{1/2}z^u_{co}$	$\frac{1}{2}\varepsilon\tau_{\pm}x^4(x^1)^2$

REFERENCES

S. BENENTI AND M. FRANCAVIGLIA (1980), *The theory of separability of the Hamilton–Jacobi equation and its applications to general relativity*, in General Relativity and Gravitation, 1, A. Held, ed., Plenum, New York.

G. BLUMAN AND J. COLE (1974), *Similarity Methods for Differential Equations*, Applied Mathematical Sciences 13, Springer, New York.

M. BÖCHER (1894), *Die Reihenentwickelungen der Potentialtheorie*, Leipzig.

C. P. BOYER (1974), *The maximal kinematical invariance group for an arbitrary potential*, Helv. Phys. Acta, 47, pp. 589–605 (Neiderer (1973) was the first to demonstrate the equivalence of the harmonic oscillator and linear potential Schrödinger equations.)

_____ (1976), *Lie theory and separation of variables for the equation* $U_t+\Delta_2 U-(\alpha/x_1^2+\beta/x_2^2)U=0$, SIAM J. Math. Anal., 7, pp. 230–263.

C. P. BOYER, E. G. KALNINS AND W. MILLER, JR. (1975a), *Lie theory and separation of variables. 6. The equation* $U_t+\Delta_2 U=0$, J. Math. Phys., 16, pp. 499–511.

_____ (1975b), *Lie theory and separation of variables. 7. The Harmonic Oscillator in elliptic coordinates and Ince polynomials*, J. Math. Phys., 16, pp. 512–517.

L. P. EISENHART (1934), *Separable systems of Stäckel*, Ann. Math., 35, pp. 284–305.

_____ (1949), *Riemannian Geometry*, Princeton Univ. Press, Princeton, NJ.

M. HAUSNER AND J. SCHWARTZ (1968), *Lie Groups and Lie Algebras*, Gordon & Breach, New York.

E. G. KALNINS AND W. MILLER, JR. (1974), *Lie theory and separation of variables. 5. The equations* $iU_t+U_{xx}=0$ *and* $iU_t+U_{xx}-(c/x^2)U=0$, J. Math. Phys., 15, pp. 1728–1737.

_____ (1977), *Lie theory and the wave equation in space-time. 3. Semisubgroup coordinates*, J. Math. Phys., 18, p. 271.

_____ (1979), *Non-orthogonal separable coordinate systems for the flat 4-space Helmholtz equation*, J. Phys. A. Math. Gen., 12, pp. 1129–1147.

_____ (1981), *Killing tensors and nonorthogonal variable separation for Hamilton–Jacobi equations*, this Journal, 12, pp. 617–629.

_____ (1982a), *Separation of variables on n-dimensional Riemannian manifolds 1. The n-sphere and Euclidean n-space*. Univ. Minnesota Mathematics Report 81–158, (to be published).

_____ (1982b), *Separation of variables on n-dimensional Riemannian manifolds. 3. Conformally Euclidean spaces* C_n, Univ. Minnesota Mathematics Report 81–171 (to be published).

_____ (1983), *Intrinsic characterisation of variable separation for the partial differential equations of mechanics*, in Proc. IUTAM-ISIMM Symposium on Modern Developments in Analytical Mechanics, S. Benenti, M. Francaviglia and A. Lichnerowicz, eds., Academia delle Scienze di Torino, Torino.

_____ (1984), *R-separation of variables for the time-dependent Hamilton–Jacobi equation*, Univ. Waikato Mathematics Report 123, Hamilton, New Zealand (to be published). Although not cited, this paper is a generalisation of the work carried out for (*) where Δ_m is replaced by the Laplace–Beltrami operator of an arbitrary Riemannian space. In particular the authors show that if the Laplace Beltrami operator is that of a space of constant curvature, *R*-separation cannot occur.

W. MILLER, JR. (1977), *Symmetry and Separation of Variables*, Addison-Wesley, Reading, MA.

U. NEIDERER (1973), *The maximal kinematical group of the harmonic oscillator*, Helv. Phys. Acta., 46, pp. 191–200.

G.J. REID (1984), *Variable separation for heat and Schrödinger equations*, Ph. D. Thesis, Univ. Waikato, Hamilton, New Zealand.

H. P. ROBERTSON (1927), *Bemerkung über separierbare Systeme in der Wellenmechanik*, Math. Annal., 98, pp. 749–752.

V. N. SHAPOVALOV (1979), *Stäckel spaces*, Sibirskii Mat. Zh., 20, (1979), 1117–1130; Siberian Math. J., 20, (1980), pp. 790–800.

V. N. SHAPOVALOV AND N. B. SUKHOMLIN (1974), *Separation of variables in tehnonstationary Schrödinger equation*, Izv. Vyss. Ucheb. Zaved., Fizika, 12, pp. 100–105; Soviet Physics, 17 (12), (1976), pp. 1718–1722.

J. D. TALMAN (1968), *Special Functions (A Group-Theoretic Approach)*, Benjamin, New York.

T. Y. THOMAS (1946), *The fundamental theorem on quadratic first integrals*, Proc. Nat. Acad Sci. 32, pp. 10–15.

D. V. WIDDER (1975), *The Heat Equation*, Academic Press, New York.

SIAM J. MATH. ANAL.
Vol. 17, No. 3, May 1986

011

AN INTEGRAL TRANSFORM INVOLVING HEUN FUNCTIONS AND A RELATED EIGENVALUE PROBLEM*

G. VALENT†

Abstract. An integral transform involving Heun functions is obtained. When combined with the explicit solutions given by Carlitz new closed integral representations are obtained for some Heun functions. As an application we solve an eigenvalue problem related to birth and death processes obtaining the exact spectrum and eigenfunctions. A direct proof of their orthogonality and completeness is given.

1. Introduction. Heun's differential equation [1] is the most general second order differential equation with four regular singular points located at $z=0,\ 1,\ 1/k^2,\ \infty$ where we take the real parameter k^2 in the domain $\mathcal{D}_0=\{k^2 \mid 0\leqq k^2\leqq k_0^2<1\}$ for some fixed k_0.

Heun's equation is given by

$$\frac{d^2y}{dz^2}+\left(\frac{\gamma}{z}-\frac{\delta}{1-z}-\frac{\varepsilon k^2}{1-k^2z}\right)\frac{dy}{dz}+\frac{\alpha\beta k^2z+s}{z(1-z)(1-k^2z)}y=0, \tag{1.1}$$

$$\alpha+\beta=\gamma+\delta+\varepsilon-1, \tag{1.2}$$

where s is the so-called "accessory parameter".

Most of what is known about solutions of this equation is summarized in [2, p. 57]. The most important result is the existence of expansions of Heun's functions in terms of hypergeometric functions. The coefficients of such expansions obey a three term recurrence formula which makes them hard to use. Furthermore it does not seem possible to obtain explicitly the eigenvalues from such expansions.

The solutions of (1.1) which may be considered as the elliptic generalization of the hypergeometric function (in the same sense as Jacobi elliptic functions generalize the trigonometric ones) are of interest in many physical problems. Nevertheless little effort has been made to obtain explicit integral representations for them. Since the fifties, as far as we know, the main progress has been due to Carlitz [3] who obtained a finite set of nontrivial exact solutions for (1.1). (By nontrivial, we mean solutions for arbitrary values of the accessory parameter s.) It is interesting to note that he found these solutions in a study of orthogonal polynomials for which Heun's functions appear as generating functions and not in a direct analysis of (1.1).

In this article we shall give an integral transform relating two Heun's functions with different parameters. This comes about if one looks for a solution of (1.1) as a Mellin transform (in the sense of [4, p. 195]) whose kernel is itself a Heun function. Unexpectedly this transform is successful, and when combined with the explicit solutions given by Carlitz, it leads to a finite set of new integral representations for Heun's functions.

As an application we discuss thoroughly an eigenvalue problem which is relevant to the study of quadratic birth and death processes (see [5] for an introduction). Upon use of a previously derived integral representation we obtain the exact eigenvalues and eigenfunctions.

*Received by the editors January 14, 1983, and in revised form June 6, 1984.

†Laboratoire de Physique Théorique et Hautes Energies, Université Paris VII, 75251 Paris Cedex, France. The Laboratory is associated with Centre National de la Recherche Scientifique.

Contrary to the case where $k^2=0$, for nonvanishing k^2 the eigenfunctions *are no longer polynomials* with respect to the variable. Nevertheless it is possible to give a direct proof of the orthogonality of the eigenfunctions, to compute their norm and to prove their completeness in the space of square integrable functions with the weight $w(x)=(1-x)^{-1}$.

The eigenvalues display a continuous dependence for $k^2\in\mathscr{D}_0$ while for $k^2\to 1$ they all vanish, a fact related to a deep change in the nature of the spectrum which becomes continuous.

2. Connection relations for Heun functions. We shall denote a Heun function by the symbol

$$H(\alpha,\beta;\gamma,\delta,\varepsilon;k^2,s;z)\equiv H(P;k^2,s;z),$$
$$P=\{\alpha,\beta;\gamma,\delta,\varepsilon\}.$$

By a Heun function, we mean the solution of (1.1) which is holomorphic in a finite neighbourhood of $z=0$. For $\mathrm{Re}\,\gamma>0$, using the Wronskian, we readily prove that this solution is *unique*. In all that follows, we take $\mathrm{Re}\,\gamma>0$, and normalize H by

$$H(P;k^2,s;0)=1. \tag{2.1}$$

To make things more precise, we shall suppose that H is analytic in some domain $D=\{z\,||z|\leqq R\}$ with fixed (and at the moment unknown) $R>0$. R will be prescribed later.

We define a transformation of the parameters P by

$$P'=T_\alpha P\Leftrightarrow\begin{cases}\gamma'=\alpha,\\ \alpha'=\gamma,\\ \delta'=\delta+\gamma-\alpha,\\ \beta'=\beta,\\ \varepsilon'=\varepsilon+\gamma-\alpha.\end{cases} \tag{2.2}$$

This transformation is such that relation (1.2) still holds for the new parameters.

We consider the function

$$G(z)=\int_0^1 dt\,t^{c-1}(1-t)^{\gamma-c-1}H(P';k^2,s;zt)$$
$$\equiv\int_0^1 dt\,U(t)H(P';k^2,s;zt),$$

with the restrictions

$$\mathrm{Re}\,\gamma>\mathrm{Re}\,c>0. \tag{2.3}$$

Let us define the differential operator

$$\mathscr{L}_z(P)=z(1-z)(1-k^2z)\frac{d^2}{dz^2}$$
$$+\left[\gamma(1-z)(1-k^2z)-\delta z(1-k^2z)-\varepsilon k^2z(1-z)\right]\frac{d}{dz}+\alpha\beta k^2z+s.$$

We shall prove that $G(z)$ is holomorphic for $z\in D$ and is annihilated by $\mathscr{L}_z(P)$. This will imply that $G(z)$ is a Heun function with parameters P.

We divide the proof into several lemmas.

LEMMA 1. *$G(z)$ is holomorphic for $z\in D$.*

Proof. $H(P';k^2,s;zt)$ and $(\partial/\partial z)H(P';k^2,s;zt)$ are holomorphic with respect to z uniformly for $t\in[0,1]$. One has the obvious inequality

$$\left|t^{c-\alpha}(1-t)^{\gamma-c-1}\frac{\partial}{\partial z}H(P';k^2,s;zt)\right|\leqq t^{\mathrm{Re}(c-1)}(1-t)^{\mathrm{Re}(\gamma-c-1)}\left|\frac{\partial}{\partial z}H(P';k^2,s;zt_0)\right|$$

for some $t_0\in[0,1]$ and the left-hand member is integrable over $t\in[0,1]$ because of (2.3). This shows the holomorphy of $G(z)$ for $z\in D$ and the legitimacy of bringing the derivatives with respect to z inside the integral over t.

Next we prove:

LEMMA 2. $\mathscr{L}_z(P)G(z)=0, \qquad z\in D.$

Proof. Using Lemma 1, a straightforward but lengthy computation, we obtain

$$\mathscr{L}_z(P)G(z)=\int_0^1 dtU(t)M_vH(P';k^2,s;v)+A/z+B\cdot(k^2z) \tag{2.4}$$

where

$$v=zt,$$

$$M_v=v(1-v)(1-k^2v)\frac{d^2}{dv^2}+\{c-[\gamma+\delta+(\gamma+\varepsilon)k^2]v+dv^2\}\frac{d}{dv}+s+ev,$$

$$A=\int_0^1 dtU(t)\left[t(1-t)\frac{\partial^2}{\partial t^2}+(\gamma t-c)\frac{\partial}{\partial t}\right]H(P';k^2,s;v),$$

$$B=\int_0^1 dtU(t)\left\{t^2(1-t)\frac{\partial^2}{\partial t^2}+t\left(1+\alpha+\beta-\frac{dt}{k^2}\right)\frac{\partial}{\partial t}+\alpha\beta-e\cdot\frac{t}{k^2}\right\}H(P';k^2,s;v).$$

At that stage the constants c,d,e are free parameters to be adjusted later on.

As is apparent from (2.4), it is crucial for this integral transform to be successful that the coefficients A and B vanish.

Let us begin with A. Integrating by parts, we find that

$$A=-t^c(1-t)^{\gamma-c}\frac{\partial}{\partial t}H(P';k^2,s;zt)\Big|_{t=0}^{t=1} \tag{2.5}$$

since

$$U(t)=t^{c-1}(1-t)^{\gamma-c-1}.$$

On account of (2.3) and the holomorphy of Heun function, the right side of (2.5) vanishes.

We now consider B. We choose the constants c,d,e as

$$c=\alpha, \quad d=(\beta+\gamma+1)k^2, \quad e=\beta\gamma k^2$$

and integrate by parts twice to obtain

$$B=t^2(1-t)U(t)\frac{\partial}{\partial t}H(P';k^2,s;zt)+\beta t(1-t)U(t)H(P';k^2,s;zt)\Big|_{t=0}^{t=1}.$$

These terms vanish in view of (2.3) which now reads

$$\operatorname{Re}\gamma>\operatorname{Re}\alpha>0.$$

Using the values of c,d,e, we note that M_v is nothing but $\mathscr{L}_v(P')$ which precisely annihilates $H(P';k^2,s;v)$. This constitutes the proof of Lemma 2.

An immediate consequence is:

THEOREM 1. *If* 1) $\operatorname{Re}\gamma>\operatorname{Re}\alpha>0$, *and* 2) $P'=T_\alpha P$ *given by* (2.2), *then for* $z\in D$ *we have*

$$(2.6)\qquad H(P;k^2,s;z)=\frac{\Gamma(\gamma)}{\Gamma(\alpha)\Gamma(\gamma-\alpha)}\int_0^1 dt t^{\alpha-1}(1-t)^{\gamma-\alpha-1}H(P';k^2,s;zt).$$

Proof. $G(z)$ is holomorphic for $z\in D$ and is annihilated by $\mathscr{L}_z(P)$. We know that there is a unique solution to this problem which is $H(P;k^2,s;z)$ and hence $G(z)$ is related to it up to a multiplicative constant which is obtained by letting $z\to 0$ and using (2.1).

Analytic continuation will extend Theorem 1 to the complex plane with a cut along the positive real axis for $\operatorname{Re}z\geqq 1$.

Let us make some remarks on this result.

First, Erdélyi had already obtained [6] an integral equation for Heun functions. In his case, however, $P'=P$ and he deals with a kernel containing at least one hypergeometric function. His result is therefore completely different from (2.6).

Secondly, it is interesting to see what happens when $k^2\to 0$. In this case Heun functions become hypergeometric functions, and (2.6) degenerates in an integral relation due to Bateman

$${}_2F_1(a,b;c;z)=\frac{\Gamma(c)}{\Gamma(\alpha)\Gamma(c-\alpha)}\int_0^1 dt t^{\alpha-1}(1-t)^{c-\alpha-1}{}_2F_1(a,b;c;zt),\quad \operatorname{Re}c>\operatorname{Re}\alpha>0.$$

Thirdly, since $H(P';k^2,s;z)$ is a symmetric function of α and β, we can write down (2.6) with β in place of α everywhere in the right-hand member (and in that case $\operatorname{Re}\gamma>\operatorname{Re}\beta>0$).

We define

$$P'=T_\beta P\Leftrightarrow\begin{cases}\alpha'=\alpha, & \gamma'=\beta,\\ \beta'=\gamma, & \delta'=\delta+\gamma-\beta,\\ & \varepsilon'=\varepsilon+\gamma-\beta.\end{cases}$$

Combining T_α and T_β, we have

$$T_\alpha^2=T_\beta^2=\text{identity},$$
$$T_\alpha T_\beta=T_\alpha,\qquad T_\beta T_\alpha=T_\beta,$$

which shows that no new relation can be obtained by iteration.

3. Integral representations for some particular Heun functions. Let us first recall the remarkable results obtained by Carlitz [3]. We refer to the article for the derivation and we give only the results.

Consider

$$f_1=\cos\left(2\sqrt{s}\,\mathrm{sn}^{-1}(\sqrt{z}\,;k^2)\right),$$
$$f_2=\sin\left(2\sqrt{s}\,\mathrm{sn}^{-1}(\sqrt{z}\,;k^2)\right)$$

where the square root determination is taken to be positive for real positive argument.

The inverse function of the Jacobi elliptic function $\mathrm{sn}(z;k^2)$ is also multiple valued and we take the branch which vanishes for $z=0$.

One can check that f_1, f_2 are two linearly independent solutions of Heun's equations with parameters

$$\alpha=0,\quad \beta=\tfrac{1}{2},\quad \gamma=\delta=\varepsilon=\tfrac{1}{2}.$$

Putting

$$f=z^a(1-z)^b(1-k^2z)^c g,$$
$$a\left(a-\tfrac{1}{2}\right)=b\left(b-\tfrac{1}{2}\right)=c\left(c-\tfrac{1}{2}\right)=0,$$

we find, in turn, other solutions with different parameters. We list below the explicit form of Carlitz solutions normalized as in (2.1)

(3.1)

$$H\left(0,\frac{1}{2};\frac{1}{2},\frac{1}{2},\frac{1}{2};k^2,s_0;z\right)=\cos\left(2\sqrt{s_0}\,\zeta\right),$$
$$H\left(\frac{1}{2},1;\frac{3}{2},\frac{1}{2},\frac{1}{2};k^2,s_0-\frac{1+k^2}{4};z\right)=\frac{\sin\left(2\sqrt{s_0}\,\zeta\right)}{2\sqrt{s_0}\sqrt{z}},$$

$$H\left(\frac{1}{2},1;\frac{1}{2},\frac{3}{2},\frac{1}{2};k^2,s_0-\frac{1}{4};z\right)=\frac{\cos\left(2\sqrt{s_0}\,\zeta\right)}{\sqrt{1-z}},$$
$$H\left(\frac{1}{2},1;\frac{1}{2},\frac{1}{2},\frac{3}{2};k^2,s_0-\frac{k^2}{4};z\right)=\frac{\cos\left(2\sqrt{s_0}\,\zeta\right)}{\sqrt{1-k^2z}},$$

$$H\left(1,\frac{3}{2};\frac{3}{2},\frac{3}{2},\frac{1}{2};k^2,s_0-1-\frac{k^2}{4};z\right)=\frac{\sin\left(2\sqrt{s_0}\,\zeta\right)}{2\sqrt{s_0}\sqrt{z(1-z)}},$$
$$H\left(1,\frac{3}{2};\frac{3}{2},\frac{1}{2},\frac{3}{2};k^2,s_0-k^2-\frac{1}{4};z\right)=\frac{\sin\left(2\sqrt{s_0}\,\zeta\right)}{2\sqrt{s_0}\sqrt{z(1-k^2z)}},$$

$$H\left(1,\frac{3}{2};\frac{1}{2},\frac{3}{2},\frac{3}{2};k^2,s_0-\frac{1+k^2}{4};z\right)=\frac{\cos\left(2\sqrt{s_0}\,\zeta\right)}{\sqrt{(1-z)(1-k^2z)}},$$

$$H\left(\frac{3}{2},2;\frac{3}{2},\frac{3}{2},\frac{3}{2};k^2,s_0-1-k^2;z\right)=\frac{\sin\left(2\sqrt{s_0}\,\zeta\right)}{2\sqrt{s_0}\sqrt{z(1-z)(1-k^2z)}},$$

$$\zeta=\mathrm{sn}^{-1}(\sqrt{z}\,;k^2).$$

Using Theorem 1, we get four new integral representations

$$
\begin{aligned}
&H\left(\frac{1}{2},\frac{1}{2};1,0,1;k^2,s_0-\frac{k^2}{4};z\right)=\frac{1}{\pi}\int_0^1 \frac{dt}{\sqrt{t(1-t)}}\cdot\frac{\cos\left(2\sqrt{s_0}\,\mathrm{sn}^{-1}(\sqrt{zt};k^2)\right)}{\sqrt{1-k^2zt}},\\
&H\left(\frac{1}{2},\frac{1}{2};1,1,0;k^2,s_0-\frac{1}{4};z\right)=\frac{1}{\pi}\int_0^1 \frac{dt}{\sqrt{t(1-t)}}\cdot\frac{\cos\left(2\sqrt{s_0}\,\mathrm{sn}^{-1}(\sqrt{zt};k^2)\right)}{\sqrt{1-zt}},
\end{aligned}
\tag{3.2}
$$

$$
H\left(\frac{1}{2},\frac{3}{2};1,1,1;k^2,s_0-\frac{1+k^2}{4};z\right)=\frac{1}{\pi}\int_0^1 \frac{dt}{\sqrt{t(1-t)}}\cdot\frac{\cos\left(2\sqrt{s_0}\,\mathrm{sn}^{-1}(\sqrt{2t};k^2)\right)}{\sqrt{(1-zt)(1-k^2zt)}},
$$

$$
H\left(\frac{3}{2},\frac{3}{2};2,1,1;k^2,s_0-1-k^2;z\right)=\frac{2}{\pi}\int_0^1 \frac{dt\sqrt{t}}{\sqrt{1-t}}\cdot\frac{\sin\left(2\sqrt{s_0}\,\mathrm{sn}^{-1}(\sqrt{zt};k^2)\right)}{2\sqrt{s_0}\sqrt{zt(1-zt)(1-k^2zt)}}.
$$

In order to use Theorem 1, one may take $R<1$; analytic continuation extends the relations (3.2) to the whole complex plane with a cut on the real axis for $\mathrm{Re}\,z\geqq 1$. If $\alpha=\frac{1}{2}$, $\beta=1$, $\gamma=\frac{3}{2}$, $\delta=\varepsilon=\frac{1}{2}$ formula (2.6) gives an integral which can be computed and one recovers the results already given by (3.1).

Theorem 1 may generate additional formulas. For example let $\zeta=1-z$. As a function of ζ, the Heun function has the parameters

$$
P':\begin{cases}\alpha'=\alpha, & \gamma'=\delta, \quad \varepsilon'=\varepsilon, \quad k'^2=1-k^2,\\ \beta'=\beta, & \delta'=\varepsilon, \quad s'=\dfrac{s+\alpha\beta k^2}{k'^2},\end{cases}
$$

and k^2 is transformed into $-k^2/k'^2$.

Considering the solution which is holomorphic in a neighbourhood of $\zeta=0$, we can again prove Theorem 1 and obtain new integral representations.

In this example, we took $\zeta=1-z$, but any one among the 24 changes of variables given in [7, p. 577] will work. We shall not give all the related formulas which can be easily obtained.

4. Application to an eigenvalue problem. Among the integral representations obtained in (3.2), one is relevant for application to a birth and death process with quadratic transition rates (see, for instance [5] for an introduction).

One is led to the following eigenvalue problem:

$$
\left\{(1-x)\frac{d}{dx}\left[x(1-k^2x)\frac{d}{dx}\right]-\frac{k^2}{4}(1-x)-s_0\right\}y_{s_0}(x)=0 \tag{4.1}
$$

with $k^2\in\mathscr{D}_0=\{k^2\,|\,0\leqq k^2\leqq k_0^2<1\}$ (the analysis for $k^2=1$ is given in [5]). The boundary conditions are

$$
y_{s_0}(0)=1, \qquad y_{s_0}(1)=0. \tag{4.2}
$$

4.1. Eigenfunctions and eigenvectors. The first condition in (4.2) implies that the eigenfunctions are

$$
y_{s_0}(x;k^2)=H\left(\frac{1}{2},\frac{1}{2};1,0,1;k^2,s_0-\frac{k^2}{4};x\right).
$$

The other linearly independent solution is easily excluded because it has a logarithmic singularity at $x=0$.

The following lemma is essential to identify the eigenvalues.

LEMMA 3. *If $k^2\in\mathscr{D}_0$ the eigenfunctions $y_{s_0}(x;k^2)$ are continuous for $x\in[0,1]$ and one has*

$$y_{s_0}(1;k^2)=\frac{\sin\left(2\sqrt{s_0}K\right)}{\sqrt{s_0}\pi}$$

where $K(k^2)=\pi/2.\ {}_2F_1(\frac{1}{2},\frac{1}{2};1,k^2)$ is the complete elliptic integral of the first kind.

Proof. We use the integral representation of $y_{s_0}(x;k^2)$ obtained in (3.2)

$$y_{s_0}(x;k^2)=\frac{1}{\pi}\int_0^1\frac{dt}{\sqrt{t(1-t)}}\cdot\frac{\cos\left(2\sqrt{s_0}\,\mathrm{sn}^{-1}(\sqrt{xt}\,;k^2)\right)}{\sqrt{1-k^2xt}}. \tag{4.3}$$

The integrand is a continuous function of $x\in[0,1]$ for $t\in[0,1]$. Furthermore we have the bound

$$\frac{1}{\sqrt{t(1-t)}}\left|\frac{\cos\left(2\sqrt{s_0}\,\mathrm{sn}^{-1}(\sqrt{xt}\,;k^2)\right)}{\sqrt{1-k^2xt}}\right|\leqq\frac{\mathrm{ch}\left(2K\cdot\mathrm{Im}\sqrt{s_0}\right)}{\sqrt{t(1-t)(1-k^2t)}}=g(t)$$

with $g(t)$ obviously integrable for $t\in[0,1]$. Hence we get the continuity of $y_{s_0}(x;k^2)$ and

$$y_{s_0}(1;k^2)=\frac{1}{\pi}\int_0^1\frac{dt}{\sqrt{t(1-t)}}\cdot\frac{\cos\left(2\sqrt{s_0}\,\mathrm{sn}^{-1}(\sqrt{t}\,;k^2)\right)}{\sqrt{1-k^2t}}.$$

Putting $t=\mathrm{sn}^2(\theta;k^2)$, we get

$$y_{s_0}(1;k^2)=\frac{2}{\pi}\int_0^K d\theta\cos\left(2\sqrt{s_0}\theta\right)=\frac{\sin\left(2\sqrt{s_0}K\right)}{\pi\sqrt{s_0}}.$$

As a side remark the simplicity of this main result is made more transparent if in (4.3) we change the variables to

$$x=\mathrm{sn}^2(\theta;k^2),\quad xt=\mathrm{sn}^2(\psi;k^2),\quad \theta,\psi\in[0,K].$$

Then the eigenfunctions may be written

$$y_{s_0}\left(\mathrm{sn}^2(\theta;k^2);k^2\right)=\frac{2}{\pi}\int_0^\theta d\psi\frac{\mathrm{cn}(\psi;k^2)}{\sqrt{\mathrm{sn}^2(\theta;k^2)-\mathrm{sn}^2(\psi;k^2)}}\cos\left(2\sqrt{s_0}\psi\right).$$

From Lemma 3 we conclude that the eigenvalues are given by

$$(s_0)_n=\frac{n^2\pi^2}{4K^2(k^2)}=n^2\frac{K^2(0)}{K^2(k^2)},\qquad n\in\mathbb{N}^*, \tag{4.4}$$

$$\mathbb{N}^*=\{1,2,3,\cdots.\}$$

and the eigenfunctions

$$y_n(x;k^2)=\frac{1}{\pi}\int_0^1\frac{dt}{\sqrt{t(1-t)}}\cdot\frac{\cos\left((n\pi/K)\,\mathrm{sn}^{-1}(\sqrt{xt}\,;k^2)\right)}{\sqrt{1-k^2xt}}. \tag{4.5}$$

All the eigenvalues are simple.

If $k^2 \to 0$ we recover for eigenvalues n^2 with eigenfunctions ${}_2F_1(-n, n; 1, x)$ which are hypergeometric *polynomials*.

For $k^2 \in \mathscr{D}_0$ the eigenvalues exhibit a continuous dependence with respect to k^2, but if $k^2 > 0$ the eigenfunctions *are no longer polynomials* and have a complicated structure displayed by (4.5). If $k^2 \to 1$ (which lies outside $\mathscr{D}_0$) $K(k^2)$ has a logarithmic singularity and diverges. All the eigenvalues collapse to zero, an indication of a drastic change in the spectrum which becomes a continuous one (this was observed in [5]).

This is rather interesting from a differential equation theoretic point of view since we have at our disposal a soluble model where, according to the value of the parameter k^2, we can "see" the transition from a discrete spectrum to a continuous one.

One can check (4.4) using perturbation theory around $k^2 = 0$. Up to order $(k^2)^2$ the perturbation theory calculations are in agreement with (4.4).

4.2. Summation formulae for the eigenfunctions. As an interesting application of the integral representation (4.5) we have to mention the possibility of evaluating several sums involving the eigenfunctions $y_n(x; k^2)$.

Let us first give an example. We start from the series given, for instance, in [7, p. 511]

$$\mathrm{dn}(\psi; k^2) = \frac{\pi}{2K} + \frac{2\pi}{K} \sum_{n=1}^{\infty} \frac{q^n}{1+q^{2n}} \cos\left(\frac{n\pi\psi}{K}\right) = \sum_{n=0}^{\infty} \mu_n(\psi), \qquad \psi \in [0, K].$$

For $k^2 \in \mathscr{D}_0$ we have $q = e^{-\pi K'/K} < 1$ and the series is absolutely and uniformly convergent for $\psi \in [0, K]$. Furthermore it is easy to see that the series

$$\sum_{n=0}^{\infty} |\mu_n(\psi)|$$

is integrable over $[0, K]$. Hence we may multiply each term by

$$\frac{\mathrm{cn}(\psi; k^2)}{\sqrt{\mathrm{sn}^2(\theta; k^2) - \mathrm{sn}^2(\psi; k^2)}}$$

and integrate term by term from $\psi = 0$ to $\psi = \theta$. This gives

$$\frac{2}{\pi} \int_0^{\theta} d\psi \frac{\mathrm{cn}\psi\,\mathrm{dn}\psi}{\sqrt{\mathrm{sn}^2\theta - \mathrm{sn}^2\psi}} = \frac{\pi}{2K} y_0(\mathrm{sn}^2\theta) + \frac{2\pi}{K} \sum_{n=1}^{\infty} \frac{q^n}{1+q^{2n}} \cdot y_n(\mathrm{sn}^2\theta).$$

The integral on the left side equals $\pi/2$. Taking into account the relation

$$y_0(x) = {}_2F_1\left(\frac{1}{2}, \frac{1}{2}; 1; k^2 x\right),$$

we obtain the summation formula

$$\frac{2\pi}{K} \sum_{n=1}^{\infty} \frac{q^n}{1+q^{2n}} y_n(x) = 1 - \frac{\pi}{2K} {}_2F_1\left(\frac{1}{2}, \frac{1}{2}; 1; k^2 x\right). \tag{4.6}$$

For $k^2 \in \mathscr{D}_0$ the left-hand member series is uniformly convergent for $x \in [0, 1]$. The reader can check that this series defines a C^∞ function for $x \in [0, 1]$ and that it is legitimate to differentiate it term by term. Hence differentiating (4.6) with respect to x will give new relations.

Similarly, using the known Fourier series (see [7]) for $1/\mathrm{dn}\psi$ and $\mathrm{sn}^2\psi$, we find that

$$\frac{2\pi}{K}\sum_{n=1}^{\infty}(-1)^n\frac{q^n}{1+q^{2n}}\cdot y_n(x)=\frac{k'}{\sqrt{1-k^2x}}-\frac{\pi}{2K}\,{}_2F_1\left(\frac{1}{2},\frac{1}{2};1;k^2x\right),$$

$$\frac{2\pi^2}{K^2}\sum_{n=1}^{\infty}\frac{nq^n}{1-q^{2n}}\cdot y_n(x)={}_2F_1\left(-\frac{1}{2},\frac{1}{2};1;k_x^2\right)-\frac{E}{K}\,{}_2F_1\left(\frac{1}{2},\frac{1}{2};1;k^2x\right),$$

where $E=E(k^2)$ is the complete elliptic integral of the second kind.

4.3. A useful integral representations for the eigenfunctions. The integral representation (4.5) for the eigenfunctions is not suited to prove their orthogonality because it does not exhibit their vanishing at $x=1$. This means that one can extract out from them a factor $1-x$ times some other integral. This is the aim of this paragraph, which follows the remark made at the end of §3.

We go back to the differential equation for the eigenfunctions, but for the moment s_0 is not supposed to be an eigenvalue

$$\frac{d^2y}{dx^2}+\left(\frac{1}{x}-\frac{k^2}{1-k^2x}\right)\frac{dy}{dx}+\frac{s_0-(k^2/4)(1-x)}{x(1-x)(1-k^2x)}y=0. \tag{4.7}$$

The change of variable $u=1-x$ gives

$$\frac{d^2y}{du^2}-\left(\frac{1}{1-u}+\frac{\hat{k}^2}{1-\hat{k}^2u}\right)\frac{dy}{du}+\frac{s_0/k'^2+(\hat{k}^2/4)u}{u(1-u)(1-\hat{k}^2u)}\cdot y=0, \tag{4.8}$$

$$\hat{k}^2=-k^2/k'^2,\qquad k'^2=1-k^2.$$

Equation (4.8) has a first solution

$$f_1(u)=H\left(\frac{1}{2},\frac{1}{2};0,1,1;\hat{k}^2,s_0/k'^2;u\right).$$

In order to get the other solution, we put

$$y=u\cdot Y$$

and obtain

$$\frac{d^2Y}{du^2}+\left(\frac{2}{u}-\frac{1}{1-u}-\frac{\hat{k}^2}{1-\hat{k}^2u}\right)\frac{dY}{du}+\frac{s_0/k'^2-1-\hat{k}^2+(9/4)\hat{k}^2u}{u(1-u)(1-\hat{k}^2u)}\cdot Y=0.$$

Hence, we take as the second solution of (4.8):

$$f_2(u)=uH\left(\frac{3}{2},\frac{3}{2};2,1,1;\hat{k}^2,\frac{s_0}{k'^2}-1-\hat{k}^2;u\right).$$

To prove the linear independence of f_1 and f_2, we examine their Wronskian which is

$$W[f_1,f_2]=f_1\frac{df_2}{du}-f_2\frac{df_1}{du}=-W_0(1-u)^{-1}(1-\hat{k}^2u)^{-1}.$$

An easy computation in the neighbourhood of $u=0$ gives $W_0=+1$ and we conclude the linear independence of f_1 and f_2.

Since (4.7) has at most two linearly independent solutions, we must have a linear relation of the form

$$y(x;k^2)=Af_2(u)+Bf_1(u), \qquad u=1-x.$$

Now we take s_0 to be an eigenvalue: since $y_n(x;k^2)$ vanishes at $x=1$ this implies $B=0$ and we are left with

$$(4.9)\qquad y_n(x;k^2)=A_n(1-x)H\left(\frac{3}{2},\frac{3}{2};2,1,1;\hat{k}^2,\left(\frac{n\pi}{2k'/K}\right)^2-1-\hat{k}^2;1-x\right),$$

where the unknown constant

$$A_n=-\frac{d}{dx}y_n(x;k^2)\Big|_{x=1}$$

is given by the following lemma:

LEMMA 4. *$y_n'(x;k^2)$ is continuous for $x\in[0,1]$ and*

$$\lim_{x\to 1}y_n'(x;k^2)=\frac{(-1)^n}{k'^2}\cdot\frac{n\pi}{2K}\cdot\frac{1+q^{2n}}{1-q^{2n}}, \qquad n\in\mathbb{N}^*, \quad k^2\in\mathscr{D}_0.$$

Proof. We start by using the integral representation (4.5). Formally, the derivative is given by

(4.10)

$$y_n'(x;k^2)=\frac{1}{\pi}\int_0^1 dt\sqrt{\frac{t}{(1-t)(1-k^2xt)^3}}$$

$$\cdot\left\{\frac{k^2}{2}\cdot\cos\left(\frac{n\pi}{K}\mathrm{sn}^{-1}(\sqrt{xt})\right)-\frac{n\pi}{2K}\sqrt{\frac{(1-k^2xt)}{xt(1-xt)}}\cdot\sin\left(\frac{n\pi}{K}\mathrm{sn}^{-1}(\sqrt{xt})\right)\right\}.$$

Our task is to justify differentiation under the integral sign.

We first observe that

$$(1-k^2xt)^{-3/2}\cos\left(\frac{n\pi}{K}\mathrm{sn}^{-1}(\sqrt{xt})\right)$$

is continuous for $x\in[0,1]$ and $t\in[0,1]$. We have the bound

$$\left|\sqrt{\frac{t}{(1-t)(1-k^2xt)^3}}\cdot\cos\left(\frac{n\pi}{K}\mathrm{sn}^{-1}(\sqrt{xt})\right)\right|\leqq\sqrt{\frac{t}{(1-t)(1-k^2t)^3}}=g(t)$$

where $g(t)$ is integrable for $t\in[0,1]$.

For the second piece, we must consider two cases: either $x\in[0,\frac{1}{2}]$ or $x\in[\frac{1}{2},1]$.

If $x\in[0,\frac{1}{2}]$ the function

$$(1-k^2xt)^{-1}\cdot(1-xt)^{-1/2}\cdot\frac{\sin((n\pi/K)\mathrm{sn}^{-1}(\sqrt{xt}))}{\sqrt{xt}}$$

is continuous for $t\in[0,1]$ provided that it takes the value $n\pi/K$ for $xt=0$. Hence its absolute value is bounded by some positive constant C

$$\left|\frac{t^{1/2}(1-t)^{-1/2}}{(1-k^2xt)}(1-xt)^{-1/2}\frac{\sin((n\pi/K)\mathrm{sn}^{-1}(\sqrt{xt}))}{\sqrt{xt}}\right|\leqq\frac{Ct^{1/2}(1-t)^{-1/2}}{(1-k^2t)}=h(t);$$

therefore $h(t)$ is integrable for $t\in[0,1]$.

If $x \in [\frac{1}{2}, 1]$ some care is required. Using the known relation

$$\mathrm{sn}^{-1}(\sqrt{v}\,;k^2) = K - \mathrm{sn}^{-1}\left(\sqrt{\frac{1-v}{1-k^2v}}\,;k^2\right), \qquad 0 \leqq v \leqq 1,$$

we write

$$\frac{\sin\left((n\pi/K)\mathrm{sn}^{-1}(\sqrt{xt}\,;k^2)\right)}{\sqrt{1-xt}} = (-1)^{n-1}\frac{\sin\left((n\pi/K)\mathrm{sn}^{-1}\left(\sqrt{(1-xt)/(1-k^2xt)}\,;k^2\right)\right)}{\sqrt{1-xt}}$$

and if we take this function to be equal to $(-1)^{n-1}n\pi/k'K$ for $xt=1$ it will be continuous for $x \in [\frac{1}{2}, 1]$ and $t \in [0,1]$ and its absolute value will be bounded by some constant D. So we obtain

$$\left|\frac{x^{-1/2}(1-t)^{-1/2}}{(1-k^2xt)} \cdot \frac{\sin\left((n\pi/K)\mathrm{sn}^{-1}(\sqrt{xt})\right)}{\sqrt{1-xt}}\right| \leqq \sqrt{2}\,D\frac{(1-t)^{-1/2}}{(1-k^2t)} = l(t)$$

where $l(t)$ is integrable for $t \in [0,1]$.

Collecting all the pieces, we conclude that $y_n'(x;k^2)$ is indeed continuously differentiable and it is legitimate to bring the derivative inside the integral.

We take in (4.10) the limit $x=1$ and put $t = \mathrm{sn}^2(\theta;k^2)$

$$y_n'(1;k^2) = \frac{1}{\pi}\int_0^K \frac{d\theta}{dn^2\theta} \cdot \cos\left(\frac{n\pi\theta}{K}\right) - \frac{n}{K}\int_0^K d\theta \frac{\mathrm{sn}\,\theta}{\mathrm{cn}\,\theta\,\mathrm{dn}\,\theta} \cdot \sin\left(\frac{n\pi\theta}{K}\right).$$

These integrals are computed using the calculus of residues and the elementary properties of the Jacobi elliptic functions. The result is Lemma 4.

To conclude this subsection, we notice that an integral representation for $f_2(u)$ has been obtained in (3.2). Combining it with (4.9), we get

$$\begin{aligned} y_n(x;k^2) &= Z_n(1-x)^{1/2}\int_0^1 dt \frac{\sin\left((n\pi/k'K)\mathrm{sn}^{-1}\left(\sqrt{t(1-x)}\,;\hat{k}^2\right)\right)}{\sqrt{(1-t)[1-t(1-x)][1-k^2t(1-x)]}}, \\ Z_n &= \frac{(-1)^{n-1}}{\pi k'}\frac{1+q^{2n}}{1-q^{2n}}. \end{aligned} \tag{4.11}$$

When $x = \mathrm{cn}^2(\theta;\hat{k}^2)$, with $\theta \in [0, k'K]$,

$$y_n\left(\mathrm{cn}^2(\theta;\hat{k}^2);k^2\right) = 2Z_n\int_0^\theta d\psi \frac{\mathrm{sn}(\psi;\hat{k}^2)}{\sqrt{\mathrm{sn}^2(\theta;\hat{k}^2) - \mathrm{sn}^2(\psi;\hat{k}^2)}} \cdot \sin\left(\frac{n\pi\psi}{k'K}\right). \tag{4.12}$$

This relation is essential for computing scalar products involving the eigenfunctions as we shall see now.

4.4. Orthogonality and norm of the eigenvectors. The scalar product for which the differential operator (4.1) is formally self-adjoint is

$$(f,g) = \int_0^1 \frac{dx}{1-x} f(x)g(x) \tag{4.13}$$

with real valued functions f, g.

We shall prove

THEOREM 2. *For $n, p \in \mathbb{N}^*$ we have*

$$(y_n, y_p) = \rho_n \delta_{np}, \qquad \rho_n = \frac{1}{2n}\left(\frac{2K}{\pi}\right)^2 \cdot \frac{1+q^{2n}}{1-q^{2n}}.$$

Proof. First, assume $k^2 > 0$. Substituting the integral representation (4.11) in (4.13), we have a triple integral

$$(4.14) \qquad (y_n, y_p) = Z_n Z_p \int_0^1 dx \int_0^1 dt \frac{\sin\left((n\pi/k'K)/\mathrm{sn}^{-1}\left(\sqrt{t(1-x)}\,;\hat{k}^2\right)\right)}{\sqrt{(1-t)[1-t(1-x)][1-k^2t(1-x)]}}$$

$$\times \int_0^1 ds \frac{\sin\left((n\pi/k'K)\mathrm{sn}^{-1}\left(\sqrt{s(1-x)}\,;\hat{k}^2\right)\right)}{\sqrt{(1-s)[1-s(1-x)][1-k^2s(1-x)]}}$$

$$= Z_n Z_p \int_0^1 dx \int_0^1 dt \int_0^1 ds U(x;t,s).$$

As already explained in the proof of Lemma 4, the function

$$\sin\left((n\pi/k'K)\mathrm{sn}^{-1}(\sqrt{v})\right)/\sqrt{1-v}$$

is continuous for $v \in [0,1]$ provided that, for $v=1$, we define it to be $(-1)^{n-1} n\pi/k'K$. Hence it is bounded by some constant D_n and we may write

$$|U(x;t,s)| \leqq D_n D_p \left[(1-t)(1-k^2t)(1-s)(1-k^2s)\right]^{-1/2}$$

a bound which implies the absolute convergence of the integral (4.14). Let

$$x = \mathrm{cn}^2(\theta;\hat{k}^2), \quad t(1-x) = \mathrm{sn}^2(\varphi;\hat{k}^2), \quad s(1-x) = \mathrm{sn}^2(\psi;\hat{k}^2), \qquad \theta, \varphi, \psi \in [0, k'K]$$

and interchange the orders of integration,

$$(y_n, y_p) = 4 Z_n Z_p \int_0^{k'K} d\varphi \sin\left(\frac{n\pi\varphi}{k'K}\right) \int_0^{k'K} d\psi \sin\left(\frac{p\pi\psi}{k'K}\right)$$

$$\times \left[\mathrm{sn}\varphi\,\mathrm{sn}\psi \int_{\sup(\varphi,\psi)}^{k'K} d\theta \frac{2\,\mathrm{cn}\theta\,\mathrm{dn}\theta}{\mathrm{sn}\theta} \frac{1}{\sqrt{(\mathrm{sn}^2\theta - \mathrm{sn}^2\varphi)(\mathrm{sn}^2\theta - \mathrm{sn}^2\psi)}}\right].$$

All the elliptic functions involved have $\hat{k}^2$ for parameter. The term in the bracket is elementary if one takes for variable $u = \mathrm{sn}^2(\theta;\hat{k}^2)$ and is equal to

$$\ln \frac{(\mathrm{sn}\varphi \cdot \mathrm{cn}\psi + \mathrm{cn}\varphi \cdot \mathrm{sn}\psi)^2}{|\mathrm{sn}^2\varphi - \mathrm{sn}^2\psi|}.$$

Coming back to elliptic functions with parameter k^2, we note that

$$(y_n, y_p) = 4k'^2 Z_n Z_p \int_0^K d\varphi \sin\left(\frac{n\pi\varphi}{K}\right) \int_0^K d\psi \sin\left(\frac{p\pi\psi}{K}\right) \ln \frac{(\mathrm{sn}\varphi \cdot \mathrm{cn}\psi + \mathrm{cn}\varphi \cdot \mathrm{sn}\psi)^2}{|\mathrm{sn}^2\varphi - \mathrm{sn}^2\psi|}.$$

In order to simplify the logarithm, we use relation (4) [16, p. 152]

$$(4.15) \qquad \frac{\mathrm{sn}\varphi \cdot \mathrm{cn}\psi + \mathrm{cn}\varphi \cdot \mathrm{sn}\psi}{\mathrm{dn}\psi - \mathrm{dn}\varphi} = \frac{1}{k^2} \cdot \frac{1 + \mathrm{dn}(\varphi - \psi)}{\mathrm{sn}(\varphi - \psi)}.$$

Putting in (4.15) $\psi=\psi'+2iK'$ and then changing ψ' to $-\psi'$ gives another relation which, when combined with (4.15), reads:

$$\frac{(\operatorname{sn}\varphi\operatorname{cn}\psi+\operatorname{cn}\varphi\operatorname{sn}\psi)^2}{\operatorname{sn}^2\varphi-\operatorname{sn}^2\psi}=\frac{1}{k^2}\cdot\frac{1+\operatorname{dn}(\varphi-\psi)}{\operatorname{sn}(\varphi-\psi)}\cdot\frac{1-\operatorname{dn}(\varphi+\psi)}{\operatorname{sn}(\varphi+\psi)}.$$

The scalar product is then transformed to

$$(y_n,y_p)=I_{np}^{(1)}+I_{np}^{(2)},$$

$$I_{np}^{(1)}=k'^2Z_nZ_p\int_{-K}^{+K}d\varphi\exp\left(\frac{in\pi\varphi}{K}\right)\int_{-K}^{+K}d\psi\exp\left(\frac{ip\pi\psi}{K}\right)\ln\left|\frac{k^2\operatorname{sn}(\varphi-\psi)}{1+\operatorname{dn}(\varphi-\psi)}\right|,$$

$$I_{np}^{(2)}=k'^2Z_nZ_p\int_{-K}^{+K}d\varphi\exp\left(\frac{in\pi\varphi}{K}\right)\int_{-K}^{+K}d\psi\exp\left(\frac{ip\pi\psi}{K}\right)\ln\left|\frac{\operatorname{sn}(\varphi+\psi)}{1-\operatorname{dn}(\varphi+\psi)}\right|.$$

The computation now becomes possible with the variables $\varphi\pm\psi$ and gives

$$I_{np}^{(1)}=\frac{2K}{\pi}k'^2Z_nZ_p\frac{(-1)^{n+p-1}}{n+p}\int_0^{2K}dx\left[\sin\left(\frac{n\pi x}{K}\right)+\sin\left(\frac{p\pi x}{K}\right)\right]\ln\left(\frac{\operatorname{sn}x}{1+\operatorname{dn}x}\right),$$

$$I_{np}^{(2)}=\begin{cases}\dfrac{2K}{\pi}k'^2Z_nZ_p\dfrac{(-1)^{n-p}}{n-p}\displaystyle\int_0^{2K}dx\left[\sin\left(\frac{p\pi x}{K}\right)-\sin\left(\frac{n\pi x}{K}\right)\right]\ln\left(\frac{\operatorname{sn}x}{1-\operatorname{dn}x}\right), & n\neq p,\\ 4Kk'^2Z_n^2\displaystyle\int_0^K dx\cos\left(\frac{n\pi x}{K}\right)\ln\left(\frac{\operatorname{sn}x}{1-\operatorname{dn}x}\right), & n=p.\end{cases}$$

The integrals of the form

$$\int_0^{2K}dx\ln\left(\frac{\operatorname{sn}x}{1\pm\operatorname{dn}x}\right)\sin\left(\frac{n\pi x}{K}\right),\qquad n\in\mathbb{N}^*$$

all vanish because their integrand is odd with respect to the point $x=K$. As a consequence, for $n\neq p$, the scalar product (y_n,y_p) vanishes.

For $n=p$, the remaining integral is computed using the calculus of residues applied to the function

$$e^{in\pi z/K}\left[\frac{\operatorname{cn}z\operatorname{dn}z}{\operatorname{sn}z}-\frac{k^2\operatorname{sn}z\operatorname{cn}z}{1-\operatorname{dn}z}\right].$$

The contour is a rectangle indented at $z=0$ and has vertices $-K$, $+K$, $+K+i\infty$, $-K+i\infty$. We obtain

$$\int_0^K dx\cos\left(\frac{n\pi x}{K}\right)\ln\left(\frac{\operatorname{sn}x}{1-\operatorname{dn}x}\right)=\frac{K}{2n}\cdot\frac{1-q^{2n}}{1+q^{2n}},\qquad n\in\mathbb{N}^*. \tag{4.16}$$

Using (4.16) and the explicit form of Z_n, we establish Theorem 2.

Recall that we assumed $k^2\in\mathscr{D}_0$, $k^2>0$. By letting $k^2\to 0$, we obtain the result

$$(y_n,y_p)=\frac{1}{2n}\delta_{np}.$$

Hence Theorem 2 is valid for $k^2\in\mathscr{D}_0$.

4.5. Completeness of the eigenfunctions. Let us begin with some definitions. We shall denote by L^2_W the linear vector space of square integrable functions with respect to the weight:

$$W(x)=(1-x)^{-1}, \qquad \|f\|=\left(\int_0^1 dx W(x) f^2(x)\right)^{1/2};$$

only real valued functions are considered. The scalar product is taken to be

$$(f,g)=\int_0^1 dx W(x) f(x) g(x).$$

Since the Lebesgue integral is used, L^2_W is a Hilbert space [10].

In this subsection we shall denote by $\hat{y}_n$ the orthonormal eigenfunctions.

LEMMA 5. $\mathscr{R}=\{(1-x)^n, n\in\mathbb{N}^*\}$ *is dense in* L^2_W *for the* L^2_W *norm.*

Proof. This result is an extension of the Stone–Weierstrass density theorem. It has been proved in a very general framework in [17] (see remark 1, p. 725 of this reference).

THEOREM 3. $\{\hat{y}_n, n\in\mathbb{N}^*\}$ *is a complete orthonormal basis in* L^2_W.

Proof. We consider the set of functions

$$h_r(\psi)=(\mathrm{sn}\psi)^{2r-1}\mathrm{cn}\psi\,\mathrm{dn}\psi, \qquad r\in\mathbb{N}^*$$

where all elliptic functions have for parameter $\hat{k}^2$. All these functions are odd, vanish at $\psi=\pm k'K$ and have a real period $2k'K$. Furthermore, they are C^∞. Hence their Fourier series converge uniformly for $\psi\in[-k'K,k'K]$:

$$\lim_{N\to\infty}\sup_{\psi\in[0,k'K]}\left|h_r(\psi)-\sum_{n=1}^{N}\xi_n(s)\sin\left(\frac{n\pi\psi}{k'K}\right)\right|=0.$$

The coefficients $\xi_n(r)$ are given by

$$\xi_n(r)=\frac{1}{k'K}\int_{-k'K}^{k'K} d\psi h_r(4)\sin\left(\frac{n\pi\psi}{k'K}\right). \tag{4.17}$$

Integrating by parts twice in (4.17), we find for ξ_n, the bound

$$|\xi_n|\leqq\frac{C^2}{n^2}, \qquad n\in\mathbb{N}^*, \quad \text{real } C. \tag{4.18}$$

From this the absolute convergence of the Fourier series follows. Hence we may multiply each term by $\mathrm{sn}\psi/\sqrt{\mathrm{sn}^2\theta-\mathrm{sn}^2\psi}$ and integrate term by term for $\psi\in[0,\theta]$. Using (4.12), we get

$$\sum_{n=1}^{\infty}\frac{\xi_n(r)\sqrt{\rho_n}}{2Z_n}\hat{y}_n(\mathrm{cn}^2\theta)=\int_0^\theta d\psi\frac{(\mathrm{sn}\psi)^{2r-1}\mathrm{sn}\psi\,\mathrm{cn}\psi\,\mathrm{dn}\psi}{\sqrt{\mathrm{sn}^2\theta-\mathrm{sn}^2\psi}}=C_r(\mathrm{sn}^2\theta)^r \quad r\in\mathbb{N}^*, \tag{4.19}$$

$\theta\in[-k'K,k'K]$, with pointwise convergence, where the C_r are known constants and ρ_n is given in Theorem 2. But we have the bound for sufficiently large n (use 4.5 for the eigenfunctions)

$$\left|\frac{\xi_n(r)\sqrt{\rho_n}}{2Z_n}\hat{y}_n(\mathrm{cn}^2\theta)\right|\leqq 2\pi k' y_0(\mathrm{cn}^2\theta).|\xi_n(r)|,$$

Since this upper bound is a convergent series (use (4.18)), we conclude that the series in (4.19) converges absolutely and uniformly for $\theta\in[-k'K,k'K]$. Then we make the change of variable $x=\mathrm{cn}^2(\theta;\hat{k}^2)$. For $\theta\in[0,k'K]$ and $x\in[0,1]$, it is continuous and one to one. This implies

$$\lim_{N\to\infty}\sup_{x\in[0,1]}\left|c_r(1-x)^r-S_N(x;r)\right|=0, \tag{4.20}$$

$$S_N(x;r)=\sum_{n=1}^{\infty}\frac{\xi_n(r)\sqrt{\rho_n}}{2Z_n}\hat{y}_n(x).$$

Hence $\{\hat{y}_n, n\in\mathbb{N}^*\}$ spans $\mathscr{B}$ with respect to the uniform convergence norm. This density property remains true for the L^2_W norm because of the bound

$$\int_0^1\frac{dx}{1-x}\left|c_r(1-x)^r-S_N(x;r)\right|^2\leqq\sup_{x\in[0,1-\eta]}\left|c_r(1-x)^r-S_N(x;r)\right|^2\cdot\ln\left(\frac{1}{\eta}\right)+\int_{1-\eta}^1\frac{dx}{1-x}\left|c_r(1-x)^r-S_N(x;r)\right|^2. \tag{4.21}$$

Since the function $c_r(1-x)^r-S_N(x;r)$ belongs to L^2_W for any value of N, the right side integral in (4.21) is absolutely continuous for $\eta\in[0,1]$. Therefore we can choose η such that this integral be as small as $\varepsilon/2$, independently of N.

Relation (4.20) shows that the other term in the right side of (4.21) can be made as small as $\varepsilon/2$ for some choice of N.

Hence the density of $\{\hat{y}_n, n\in\mathbb{N}^*\}$ in $\mathscr{B}$ with respect to the L^2_W norm is obtained. But for this norm $\mathscr{B}$ is dense in L^2_W [11]; therefore $\{\hat{y}_n, n\in\mathbb{N}^*\}$ is dense in L^2_W.

Since in a Hilbert space a spanning orthonormal set is automatically complete, Theorem 3 follows.

As a side remark we note that L^2_W has a remarkably rich structure of constructible orthogonal bases, since *for any value* $k^2\in\mathscr{D}_0$ the eigenfunctions $\{\hat{y}_n\}$ form a complete basis in L^2_W. Quite generally we may write

$$\hat{y}_n(x;k_1^2)=\sum_{r=1}^{\infty}e_{nr}(k_1^2,k_2^2)\,\hat{y}_r(x;k_2^2),\qquad k_1^2,k_2^2\in\mathscr{D}_0$$

from which, for $k_2^2=0$ we get

$$\hat{y}_n(x;k_1^2)=\sum_{r=1}^{\infty}\frac{e_{nr}(k_1^2,0)}{\sqrt{2r}}\,{}_2F_1(-r,r;1;x)$$

a series already given by Erdélyi [13, formula 9.1]. Using the integral representations for $\hat{y}_n(x;k_1^2)$, we may work out the coefficients e_{nr} appearing in this expansion: a nontrivial exercise !...

5. Conclusion. We have obtained an integral transform relating Heun functions with different sets of parameters. From this result and Carlitz explicit solutions, we have obtained new closed integral representations for some Heun functions.

Using this transformation, we have been able to solve an eigenvalue problem related to a birth and death process, obtaining the exact spectrum and eigenfunctions in a purely differential equation theoretic framework. The integral representations obtained are sufficient to give a direct proof of their orthogonality, to allow the computation of their norm and to prove their completeness in the Hilbert space L^2_W.

Let us observe that the spectrum obtained in this differential equation approach to birth and death processes is *identical* to that obtained by Stieltjes who directly solved the forward Kolmogorov equation using continued fraction techniques developed by himself (the continued fraction computation is given in [14] and its relevance to birth and death processes is fully discussed in [15]).

The identity of the eigenvalues in these two completely different approaches strongly suggests some underlying equivalence which may be extremely fruitful to develop both fields: continued fractions on one hand and new transcendental functions defined by differential equations on the other hand.

REFERENCES

[1] K. HEUN, Math. Ann., 33 (1889), p. 161ff.
[2] A. ERDÉLYI et al., *Higher Transcendental Functions*, Vol. III, McGraw-Hill, New York, 1955.
[3] L. CARLITZ, *Orthogonal polynomials related to elliptic functions*, Duke Math. J., 27 (1960), pp. 443–459.
[4] E. L. INCE, *Ordinary Differential Equations*, Dover, New York, 1956.
[5] B. ROEHNER AND G. VALENT, *Solving the birth and death processes with quadratic asymptotically symmetric transition rates*, SIAM J. Appl. Math., 42 (1982), p. 1020.
[6] A. ERDÉLYI, *Integral equations for Heun functions*, Quart. J. Math. Oxford, 13 (1942), pp. 107–120.
[7] E. T. WHITTAKER AND G. N. WATSON, *A Course of Modern Analysis*, Cambridge Univ. Press, Cambridge, 1965.
[8] A. ERDÉLYI, *Higher Transcendental Functions*, Vol. II, McGraw-Hill, New York, 1953.
[9] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators*, Vol. II, Interscience, New York, 1963.
[10] N. I. AKHIEZER AND I. M. GLAZMAN, *Theory of Linear Operators in Hilbert Space*, Vol. I, II, Frederick Ungar, New York, 1963.
[11] S. FOMINE AND A. KOLMOGOROV, *Introductory Real Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1970.
[12] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1966.
[13] A. ERDÉLYI, *Certain expansions of solutions of the Heun equation*, Quart. J. Math. (Oxford series), 15 (1944), pp. 62–69.
[14] T. J. STIELTJES, *Oeuvres complètes*, Vol. II, Paris, 1918.
[15] B. ROEHNER, Doctoral thesis, Paris, 1982.
[16] P. APPELL AND E. LACOUR, *Principes de la théorie des fonctions elliptiques*, Paris, 1922.
[17] I. E. SEGAL, *Abstract probability spaces and a theorem of Kolmogoroff*, Amer. J. Math., 76 (1954), pp. 721–732.

SIAM J. MATH. ANAL.
Vol. 17, No. 3, May 1986

012

DISCRETIZED FRACTIONAL CALCULUS*

CH. LUBICH†

Abstract. For the numerical approximation of fractional integrals

$$I^\alpha f(x)=\frac{1}{\Gamma(\alpha)}\int_0^x (x-s)^{\alpha-1} f(s)\,ds \qquad (x\geqq 0)$$

with $f(x)=x^{\beta-1}g(x)$, g smooth, we study convolution quadratures. Here approximations to $I^\alpha f(x)$ on the grid $x=0,h,2h,\cdots,Nh$ are obtained from a discrete convolution with the values of f on the same grid. With the appropriate definitions, it is shown that such a method is convergent of order p if and only if it is stable and consistent of order p. We introduce fractional linear multistep methods: The αth power of a pth order linear multistep method gives a pth order convolution quadrature for the approximation of I^α. The paper closes with numerical examples and applications to Abel integral equations, to diffusion problems and to the computation of special functions.

AMS (MOS) subject classifications. Primary 26A33, 41A55, 65D25; secondary 65D20, 65R20

1. Introduction. Fractional calculus is an area having a long history whose infancy dates back to the beginnings of classical calculus, and it is an area having interesting applications. The numerical approximation of the objects of classical calculus, i.e., integrals and derivatives, has for a long time been a standard topic in numerical analysis. However, the state of the art is far less advanced for fractional integrals. Hopefully, the present work contributes to narrow this gap.

Very readable introductions to fractional calculus are given by Lavoie, Osler and Tremblay [12] and by Riesz [19]. See also the book of Oldham and Spanier [18] which contains many references and applications from different areas such as special functions of mathematical physics and diffusion equations. For easy reference we collect first some basic definitions and results.

We consider *Abel–Liouville integrals* of order α (often also called Riemann–Liouville integrals),

$$I^\alpha f(x)=\frac{1}{\Gamma(\alpha)}\int_0^x (x-s)^{\alpha-1} f(s)\,ds \qquad (x\geqq 0) \quad \text{for } \operatorname{Re}\alpha>0, \tag{1.1}$$

where Γ denotes Euler's gamma function.

$I^\alpha f(x)$ depends analytically on α (for fixed f and x). If f is k-times continuously differentiable on $[0,x]$, it can be continued analytically to α with negative real part via

$$I^\alpha f(x)=\frac{d^k}{dx^k} I^{\alpha+k} f(x) \quad \text{for } \operatorname{Re}\alpha>-k. \tag{1.2}$$

If $-k\leqq \operatorname{Re}\alpha<0$ and $f^{(j)}(0)=0$ for $j=0,1,\cdots,k-1$, then $y(x)=I^\alpha f(x)$ is the solution of the first-kind *Abel integral equation*

$$\frac{1}{\Gamma(-\alpha)}\int_0^x (x-s)^{-\alpha-1} y(s)\,ds=f(x) \qquad (x\geqq 0). \tag{1.3}$$

*Received by the editors October 18, 1983, and in revised form July 13, 1984.

†Institut für Mathematik und Geometrie, Universität Innsbruck, Technikerstraße 13, A-6020 Innsbruck, Austria.

For integer α, I^α is simply repeated integration or differentiation:

$$I^k f(x)=\int_0^x\int_0^{x_k}\cdots\int_0^{x_2} f(x_1)\,dx_1\,dx_2\cdots dx_k,$$
$$I^0 f(x)=f(x),$$
$$I^{-k} f(x)=\frac{d^k}{dx^k}f(x).$$

I^α is therefore often called *fractional integral* of order α and also denoted $D^{-\alpha}$, the *fractional derivative* of order $-\alpha$.

We extend the definition to functions $f(x)=x^{\beta-1}g(x)$, where g is sufficiently differentiable for $x\geqq 0$ and $\beta\neq 0,-1,-2,\cdots$ is arbitrary. The relation

$$\left(I^\alpha\frac{t^{\beta-1}}{\Gamma(\beta)}\right)(x)=\frac{x^{\alpha+\beta-1}}{\Gamma(\alpha+\beta)}\qquad(\mathrm{Re}\,\alpha>0,\ \mathrm{Re}\,\beta>0)\tag{1.4}$$

can be used as a definition for general $\alpha,\beta\in\mathbb{C}$, $\beta\neq 0,-1,-2,\cdots$. Expanding g as a Taylor series with Bernoulli remainder we see that $(I^\alpha t^{\beta-1}g)(x)$ is then well defined.

For the numerical approximation we wish to preserve two characteristic properties of I^α:

(i) the homogeneity of I^α

$$(I^\alpha f)(x)=x^\alpha(I^\alpha f(tx))(1);$$

(ii) the convolution structure of I^α

$$I^\alpha f=\frac{1}{\Gamma(\alpha)}t^{\alpha-1}*f.$$

So we consider *convolution quadratures*

$$I_h^\alpha f(x)=h^\alpha\sum_{j=0}^{n}\omega_{n-j}f(jh)+h^\alpha\sum_{j=0}^{s}w_{nj}f(jh)\qquad(x=nh)\tag{1.5}$$

where the convolution quadrature weights ω_n $(n\geqq 0)$ and the starting quadrature weights w_{nj} $(n\geqq 0, j=0,\cdots,s;\ s$ fixed) do not depend on h.

Because of the factor h^α we have then the homogeneity relation

$$(I_h^\alpha f)(x)=x^\alpha\big(I_{h/x}^\alpha f(tx)\big)(1).$$

Also the convolution structure is essentially preserved. It is violated only by the few correction terms of the starting quadrature which will be necessary for high order schemes. For the computation of the values $I_h^\alpha f(nh)$ $(n=0,\cdots,N-1)$ one needs only N evaluations of the function f and, using fast Fourier transform techniques, only $O(N\log N)$ additions and multiplications.

There remains the important question: How have the weights ω_n and w_{nj} to be chosen in order that $I_h^\alpha f(x)$ approximate $I^\alpha f(x)$ with a prescribed order $O(h^p)$? A complete answer is given in §2. After introducing the appropriate definitions we show in Theorem 2.5 that a convolution quadrature is convergent of order p if and only if it is stable and consistent of order p. This result is an extension of Dahlquist's [3] classical

theorem on linear multistep methods. An easy way of computing a convolution quadrature of order p is by using a pth order linear multistep method to the power α (Theorem 2.6), called a fractional linear multistep method. The proofs of the results in §2 constitute §3. In §4 we give some brief remarks on the implementation of fractional linear multistep methods. §5 contains numerical examples for some applications of fractional calculus: Abel's integral equation, diffusion in a half-space, special functions of mathematical physics.

We conclude this section with a remark on the notation: If a function $f(x)$ is undefined for $x=0$, we put for simplicity $f(0)=0$. The convolution of two functions $f(x)$, $g(x)$ defined on $x \geqq 0$ is denoted by

$$(f * g)(x)=\int_0^x f(x-s)g(s)\,ds \qquad (x \geqq 0).$$

Given a sequence $a=(a_n)_0^\infty$ we denote by

$$a(\zeta)=\sum_{n=0}^{\infty} a_n \zeta^n$$

its generating power series. We do not distinguish between a formal power series, a convergent power series and the analytical function with which it coincides in its disc of convergence. We refer to (a_n) as the coefficients of $a(\zeta)$.

2. Convergence of convolution quadratures; fractional linear multistep methods. To motivate the following definitions and results we consider first the case $\alpha=1$ in (1.1) and (1.5).

If a linear multistep method (ρ,σ) (where, as usual, ρ and σ denote the generating polynomials of the method, see e.g. Henrici [8]) is applied to the quadrature problem

$$y'(x)=f(x), \quad y(0)=0, \quad \text{i.e.} \quad y(x)=\int_0^x f(s)\,ds,$$

it is well known [17], [20], [15] that the resulting numerical solution can be written as a convolution quadrature (1.5) where the weights ω_n are the coefficients of

$$\omega(\zeta)=\frac{\sigma(1/\zeta)}{\rho(1/\zeta)}. \tag{2.1}$$

The convergence of a linear multistep method is determined by its stability and consistency (Dahlquist [3], [4], also e.g. in Henrici [8]). In terms of the quadrature weights ω_n, the method is stable if and only if ω_n are bounded. Consistency of order p can be expressed as

$$h\omega(e^{-h})=1+O(h^p).$$

In the following definitions we extend these concepts to arbitrary $\alpha \in \mathbb{C}$. Here $\omega=(\omega_n)_0^\infty$ is a convolution quadrature as in (1.5).

DEFINITION 2.1. A convolution quadrature ω is *stable* (for I^α) if

$$\omega_n=O(n^{\alpha-1}).$$

DEFINITION 2.2. A convolution quadrature ω is *consistent of order p* (for I^α) if

$$h^\alpha\omega(e^{-h})=1+O(h^p).$$

Here and in the following p is a positive integer.

Remark. For $\operatorname{Re}\alpha > 0$ this condition can be interpreted as

$$h^\alpha \sum_{j=0}^{\infty} \omega_j e^{-jh} = \frac{1}{\Gamma(\alpha)} \int_0^\infty t^{\alpha-1} e^{-t}\, dt + O(h^p),$$

that is, ω yields an $O(h^p)$ approximation to the integral of the exponential function on the interval $(0, \infty)$.

For the following it is convenient to introduce the notation

$$\Omega_h^\alpha f(x) = h^\alpha \sum_{j=0}^{n} \omega_{n-j} f(jh) \qquad (x = nh), \tag{2.2}$$

which is the convolution part of (1.5), and

$$E_h^\alpha = \Omega_h^\alpha - I^\alpha, \tag{2.3}$$

the convolution quadrature error.

DEFINITION 2.3. A convolution quadrature ω is *convergent of order p* (to I^α) if

$$\left(E_h^\alpha t^{\beta-1}\right)(1) = O(h^\beta) + O(h^p) \quad \text{for all } \beta \in \mathbb{C},\ \beta \neq 0, -1, -2, \cdots. \tag{2.4}$$

This definition is motivated by the following result.

THEOREM 2.4. *Let ω satisfy* (2.4). *Then we have*:

(i) *For every $\beta \neq 0, -1, -2, \cdots$ there exists a starting quadrature*

$$w_{nj} = O(n^{\alpha-1}) \qquad (n \geqq 0, j = 0, \cdots, s) \tag{2.5}$$

such that for any function

$$f(x) = x^{\beta-1} g(x), \quad g \text{ sufficiently differentiable}, \tag{2.6}$$

the approximation $I_h^\alpha f$ given by (1.5) *satisfies*

$$I_h^\alpha f(x) - I^\alpha f(x) = O(h^p) \tag{2.7}$$

uniformly for $x \in [a, b]$ with $0 < a < b < \infty$. (More precisely, let $\tilde{\beta} = \beta + k$ (k integer) such that $0 < \operatorname{Re}\tilde{\beta} \leqq 1$. Then

$$I_h^\alpha f(x) - I^\alpha f(x) = O(x^{\alpha+\tilde{\beta}-1} h^p) \text{ uniformly for bounded } x.) \tag{2.8}$$

(ii) *For every $\beta \neq 0, -1, -2, \cdots$ there exists a starting quadrature w_{nj} (which does not necessarily satisfy* (2.5)) *such that for any function* (2.6) *the approximation $I_h^\alpha f$ satisfies* (2.7) *uniformly for bounded x.*

Remarks. a) Trivially, (i) implies (2.4).

b) The weights w_{nj} are constructed such that $I_h^\alpha t^{q+\beta-1} = I^\alpha t^{q+\beta-1}$ for all integer $q \geqq 0$ with $\operatorname{Re}(q+\beta-1) \leqq p-1$ in (i) and (ii), and additionally those with $\operatorname{Re}(q+\alpha+\beta-1) < p$ in (ii).

c) More generally, for $\beta_1, \cdots, \beta_m$ a starting quadrature (2.5) can be given for functions $f(x) = \sum_{j=1}^m x^{\beta_j - 1} g_j(x)$, g_j sufficiently differentiable, such that (2.7) holds.

In the following we consider convolution quadratures ω for which

$$\omega(\zeta) = r_1(\zeta)^\alpha r_2(\zeta) \tag{2.9}$$

where $r_i(\zeta)$ are rational functions.

We can now give the main result of this paper.

THEOREM 2.5. *A convolution quadrature* (2.9) *is convergent of order p if and only if it is stable and consistent of order p.*

Remark. a) For the special case $\alpha = 1$, Theorem 2.5 reduces in essence to Dahlquist's convergence theorem for linear multistep methods [3], [4].

b) As the proof shows, condition (2.9) can be considerably relaxed. However, the class (2.9) is probably large enough for all practical applications.

For $\alpha = k$ a positive integer, $I^k f = II \cdots I f$ (k-times) is simply the repeated integral of f. If we take $I_h f$ to be the solution of a linear multistep method (ρ, σ) applied to $y' = f$, $y(0) = 0$ (so that $y = I f$), then the repeated method $I_h^k f = I_h \cdots I_h f$ can be rewritten as a convolution quadrature (1.5) where the weights are the coefficients of the power series $\omega(\zeta)^k$, with $\omega(\zeta)$ given by (2.1). This can be interpreted as the kth power of the multistep method. We remark that squaring linear multistep methods ($k = 2$) has been used in the literature, see e.g. Dahlquist [5] and Jeltsch [10]. The following theorem shows that one can also take fractional powers of linear multistep methods. This result is a corollary of Theorem 2.5. It provides a simple means for constructing convolution quadratures for arbitrary $\alpha \in \mathbb{C}$.

THEOREM 2.6 (fractional linear multistep methods). *Let (ρ, σ) denote an implicit linear multistep method which is stable and consistent of order p. Assume that the zeros of $\sigma(\zeta)$ have absolute value less than* 1. *Let $\omega(\zeta)$, given by* (2.1), *denote the generating power series of the corresponding convolution quadrature ω. Define $\omega^\alpha = (\omega_n^{(\alpha)})_0^\infty$ by*

$$\omega^\alpha(\zeta) = \omega(\zeta)^\alpha. \tag{2.10}$$

Then the convolution quadrature ω^α is convergent of order p (to I^α).

We conclude this section with some examples.

Example 2.7. The *fractional Euler method*, $\omega^\alpha(\zeta) = (1-\zeta)^{-\alpha}$, is of historical interest. The method reads

$$I_h^\alpha f(x) = h^\alpha \sum_{0 \le jh \le x} (-1)^j \binom{-\alpha}{j} f(x - jh). \tag{2.11}$$

For $\alpha = -k$ ($k = 1, 2, 3, \cdots$) this is just the kth backward difference quotient. Starting from this observation, Liouville [14, p. 107] had already introduced fractional derivatives by a formula similar to (2.11). Grünwald [7] and Letnikov [13] have shown that (2.11) converges to the Abel–Liouville integral $I^\alpha f(x)$ (for $\operatorname{Re}\alpha > 0$). Their proof (cf. [12, p. 248], [18, p. 51]), however, does not reveal the fact that the method yields an $O(h)$-approximation.

Example 2.8. The $(p+1)$-point backward difference formula (BDF), see e.g. Henrici [8, §5.1-4], is of order p and satisfies for $p \le 6$ the assumptions of Theorem 2.6. The *fractional* BDF *methods* given in Table 1 are therefore convergent of order p. For $\alpha = -1$ the method reduces to the usual $(p+1)$-point backward difference quotient.

TABLE 1

Generating functions for $(\mathrm{BDF}p)^\alpha$, $1 \le p \le 6$.

p	$\omega^\alpha(\zeta)$
1	$(1-\zeta)^{-\alpha}$
2	$(3/2 - 2\zeta + 1/2\zeta^2)^{-\alpha}$
3	$(11/6 - 3\zeta + 3/2\zeta^2 - 1/3\zeta^3)^{-\alpha}$
4	$(25/12 - 4\zeta + 4\zeta^2 - 4/3\zeta^3 + 1/4\zeta^4)^{-\alpha}$
5	$(137/60 - 5\zeta + 5\zeta^2 - 10/3\zeta^3 + 5/4\zeta^4 - 1/5\zeta^5)^{-\alpha}$
6	$(147/60 - 6\zeta + 15/2\zeta^2 - 20/3\zeta^3 + 15/4\zeta^4 - 6/5\zeta^5 + 1/6\zeta^6)^{-\alpha}$

Example 2.9. The *fractional trapezoidal rule*, $\omega^\alpha(\zeta)=(\frac{1}{2}(1+\zeta)/(1-\zeta))^\alpha$, is convergent of order 2 if $\operatorname{Re}\alpha\geqq 0$. Since the numerator has a zero on the unit circle, the method is not stable for $\operatorname{Re}\alpha<0$ (see (3.9), (3.10)).

Example 2.10. The following class of methods can be interpreted as generalized *Newton–Gregory formulas*.

Let γ_i denote the coefficients of

$$\sum_{i=0}^{\infty}\gamma_i(1-\zeta)^i=\left(\frac{\ln\zeta}{\zeta-1}\right)^{-\alpha}$$

(see Lemma 3.2), and put

$$\omega^\alpha(\zeta)=(1-\zeta)^{-\alpha}\left[\gamma_0+\gamma_1(1-\zeta)+\cdots+\gamma_{p-1}(1-\zeta)^{p-1}\right].$$

Then ω^α is convergent of order p (to I^α). For $\alpha=1$ this method reduces to the pth order Newton–Gregory formula (i.e. implicit Adams method), for $\alpha=-1$ to the $(p+1)$-point backward difference quotient.

3. Proofs. We give first the proof of the central result, Theorem 2.5, and of its corollary, Theorem 2.6, and finally the proof of Theorem 2.4. We begin with some preparations.

Preparations. We shall repeatedly make use of the following asymptotic expansion for binomial coefficients (cf. [6, p. 47])

$$(3.1)\quad (-1)^n\binom{-\alpha}{n}=\frac{n^{\alpha-1}}{\Gamma(\alpha)}\left[1+a_1n^{-1}+a_2n^{-2}+\cdots+a_{N-1}n^{-(N-1)}+O(n^{-N})\right]$$

where the coefficients a_j depend analytically on α. $\Omega_h^\alpha f(x)$, introduced in (2.2), can be extended to

$$\Omega_h^\alpha f(x)=h^\alpha\sum_{0\leqq jh\leqq x}\omega_j f(x-jh)\qquad(x\geqq 0),$$

which is the convolution of the sequence $h^\alpha\omega$ with f. Therefore Ω_h^α commutes with convolution

$$\Omega_h^\alpha(f*g)=(\Omega_h^\alpha f)*g,$$

if f is continuous and g is locally integrable.

This property is often shared by I^α:

$$I^\alpha(f*g)=(I^\alpha f)*g$$

which holds for locally integrable g and continuous f if $\operatorname{Re}\alpha>0$, and also for f with $f^{(j)}(0)=0$ $(j=0,\cdots,k-1)$ if $\operatorname{Re}\alpha>-k$. In this case also the convolution quadrature error $E_h^\alpha=\Omega_h^\alpha-I^\alpha$ satisfies

$$(3.2)\qquad E_h^\alpha(f*g)=(E_h^\alpha f)*g.$$

The proof of the above statements is easy and therefore omitted.

The homogeneity of I^α and $t^{\beta-1}$ yields

$$(3.3)\qquad \left(E_h^\alpha t^{\beta-1}\right)(x)=x^{\alpha+\beta-1}\left(E_{h/x}^\alpha t^{\beta-1}\right)(1).$$

Formulas (3.1)–(3.3) and an analytic continuation argument will be the essential tools in the proof of Theorem 2.5.

Proof of Theorem 2.5. We break the proof into several steps which are formulated as lemmas.

LEMMA 3.1. *If* $(E_h^\alpha t^{k-1})(1)=O(h^k)+O(h^p)$ *for* $k=1,2,3,\cdots$, *then* ω *is consistent of order* p.

In particular, convergence of order p *implies consistency of order* p.

Proof. We look first at the quadrature error for e^{t-x} (as a function of t) on the interval $[0,x]$,

$$e_h(x)=\left(E_h^\alpha e^{t-x}\right)(x)=h^\alpha \sum_{0\leqq jh\leqq x} \omega_j e^{-jh}-\left(I^\alpha e^{t-x}\right)(x).$$

As $x\to\infty$, the first expression of the difference tends to $h^\alpha\omega(e^{-h})$, and

$$\left(I^\alpha e^{t-x}\right)(x)\to 1 \qquad (x\to\infty).$$

(For $\operatorname{Re}\alpha>0$ this is immediate from the definition of Euler's gamma function. For $\operatorname{Re}\alpha\leqq 0$ it follows in the same way as in the derivation of (3.5) below, with E_h^α replaced by I^α). So we have

$$e_h(\infty)=h^\alpha\omega(e^{-h})-1. \tag{3.4}$$

We expand e^{t-x} at $t=0$,

$$e^{t-x}=\sum_{k=0}^{q}\frac{t^k}{k!}e^{-x}+\frac{1}{q!}\left(\tau^q * e^{\tau-x}\right)(t),$$

with $q+1\geqq\max\{p,p-\operatorname{Re}\alpha\}$. We write

$$e_h(x)=e_h^1(x)+e_h^2(x)$$

with

$$e_h^1(x)=e^{-x}\sum_{k=0}^{q}\frac{1}{k!}\left(E_h^\alpha t^k\right)(x).$$

By (3.3), $(E_h^\alpha t^k)(x)$ has only polynomial growth as $x\to\infty$. Hence

$$e_h^1(\infty)=0.$$

By (3.2),

$$e_h^2(x)=\frac{1}{q!}E_h^\alpha\left(t^q * e^{t-x}\right)(x)=\frac{1}{q!}\left(\left(E_h^\alpha t^q\right) * e^{t-x}\right)(x)$$

$$=\frac{1}{q!}\int_0^x e^{-s}\left(E_h^\alpha t^q\right)(s)\,ds.$$

So we obtain

$$e_h(\infty)=\frac{1}{q!}\int_0^\infty e^{-s}\left(E_h^\alpha t^q\right)(s)\,ds. \tag{3.5}$$

By (3.3) and by assumption,

$$\left(E_h^\alpha t^q\right)(s)=s^{q+\alpha}\left(E_{h/s}^\alpha t^q\right)(1)=O\left(s^{q+\alpha-p}h^p\right).$$

From (3.4) and (3.5) we obtain hence

$$h^{\alpha}\omega(e^{-h})-1=O(h^{p}),$$

i.e., consistency of order p. □

Our next aim is to give in Lemma 3.2 a characterization of consistency. We may write

$$\omega(\zeta)=(1-\zeta)^{-\mu}\tilde{\omega}(\zeta)$$

where μ is chosen such that $\tilde{\omega}(\zeta)$ is holomorphic at 1 and $\tilde{\omega}(1)\neq 0$.

Consistency implies immediately $\mu=\alpha$ and $\tilde{\omega}(1)=1$. We expand $\omega(\zeta)$ at 1:

$$\omega(\zeta)=(1-\zeta)^{-\alpha}\left[c_0+c_1(1-\zeta)+c_2(1-\zeta)^2+\cdots+c_{N-1}(1-\zeta)^{N-1}+(1-\zeta)^{N}\tilde{r}(\zeta)\right] \tag{3.6}$$

where $\tilde{r}(\zeta)$ is holomorphic at 1.

We can characterize consistency in terms of the coefficients c_i.

LEMMA 3.2. *Let γ_i denote the coefficients of $\sum_0^{\infty}\gamma_i(1-\zeta)^i=(-\ln\zeta/(1-\zeta))^{-\alpha}$. Then*

ω is consistent of order p

if and only if the coefficients c_i in (3.6) *satisfy*

$$c_i=\gamma_i \quad \text{for } i=0,1,\cdots,p-1.$$

Proof. The expression

$$h^{\alpha}\omega(e^{-h})=\left(\frac{h}{1-e^{-h}}\right)^{\alpha}\tilde{\omega}(e^{-h})$$

is $1+O(h^p)$ if and only if

$$\tilde{\omega}(e^{-h})=\left(\frac{h}{1-e^{-h}}\right)^{-\alpha}+O(h^{p}),$$

which holds if and only if

$$\tilde{\omega}(\zeta)=\left(\frac{-\ln\zeta}{1-\zeta}\right)^{-\alpha}+O((1-\zeta)^{p}).$$ □

Whether the method ω is stable depends on the remainder in the expansion (3.6). We rewrite (3.6) as

$$\omega(\zeta)=(1-\zeta)^{-\alpha}\left[c_0+c_1(1-\zeta)+\cdots+c_{N-1}(1-\zeta)^{N-1}\right]+(1-\zeta)^{N}r(\zeta) \tag{3.7}$$

where $r(\zeta)=(1-\zeta)^{-\alpha}\tilde{r}(\zeta)$.

LEMMA 3.3. *ω is stable if and only if the coefficients r_n of $r(\zeta)$ in* (3.7) *satisfy*

$$r_n=O(n^{\alpha-1}). \tag{3.8}$$

Proof. It is immediate from (3.1) that (3.8) implies $\omega_n=O(n^{\alpha-1})$. Conversely, let ω be stable. Then $\omega(\zeta)$ has no singularities in the interior of the unit disc, $|\zeta|<1$, and by (2.9) can therefore be written as

$$\omega(\zeta)=u(\zeta)\prod_{j=0}^{m}(\zeta-\zeta_j)^{-\alpha_j} \tag{3.9}$$

where the ζ_j are distinct numbers of absolute value 1 (let $\zeta_0=1, \alpha_0=\alpha$), $u(\zeta)$ is holomorphic in a neighbourhood of $|\zeta|\leqq 1$, and $u(\zeta_j)\neq 0$, $\alpha_j\neq 0, -1, -2, \cdots$. Expanding $\omega(\zeta)$ at ζ_j yields (cf. partial fraction decomposition)

$$\omega(\zeta)=\sum_{j=0}^{m}(\zeta-\zeta_j)^{-\alpha_j}p_j(\zeta-\zeta_j)+q(\zeta)$$

where p_j are polynomials, $p_j(0)\neq 0$, and $q(\zeta)$ is analytic in the interior of the unit disc and sufficiently differentiable (say, k-times) on the unit circle $|\zeta|=1$, so that its coefficients are $O(n^{-k})$, (e.g. [11, p. 24]).

It is now seen from (3.1) that

$$\omega_n=O(n^{\alpha-1}) \quad \text{if and only if } \operatorname{Re}\alpha_j\leqq\operatorname{Re}\alpha \text{ for all } j. \tag{3.10}$$

Correspondingly, $r(\zeta)$ can be represented as

$$r(\zeta)=\sum_{j=0}^{m}(\zeta-\zeta_j)^{-\alpha_j}\tilde{p}_j(\zeta-\zeta_j)+\tilde{q}(\zeta)$$

with $\tilde{p}_j$ and $\tilde{q}$ as p_j and q above.

Hence (3.10) holds also with r_n instead of ω_n. This gives (3.8). □

The trivial direction of Lemma 3.3 is used in the next lemma.

LEMMA 3.4. *Convergence implies stability.*

Proof. If ω is convergent, then it is consistent by Lemma 3.1. With $N=1$ in (3.7) we have therefore

$$\omega(\zeta)=(1-\zeta)^{-\alpha}+(1-\zeta)r(\zeta).$$

We study

$$(E_h^{\alpha}1)(1)=h^{\alpha}\sum_{j=0}^{n}\omega_{n-j}-\frac{1}{\Gamma(\alpha+1)} \qquad (hn=1).$$

$\sum_{j=0}^{n}\omega_{n-j}$ is the nth coefficient of

$$\frac{\omega(\zeta)}{1-\zeta}=(1-\zeta)^{-\alpha-1}+r(\zeta).$$

By (3.1) we have

$$(E_h^{\alpha}1)(1)=h^{\alpha}\left[\frac{n^{\alpha}}{\Gamma(\alpha+1)}+O(n^{\alpha-1})\right]+h^{\alpha}r_n-\frac{1}{\Gamma(\alpha+1)}=O(h)+h^{\alpha}r_n \qquad (hn=1)$$

which is $O(h)$ only if $r_n=O(n^{\alpha-1})$. Now Lemma 3.3 completes the proof. □

It remains to show that stability and consistency imply convergence. Let us first have a closer look at the structure of the error.

LEMMA 3.5. *Let $\alpha, \beta\in\mathbb{C}$, $\beta\neq 0, -1, -2, \cdots$. If ω is stable, then the convolution quadrature error of $t^{\beta-1}$ has an asymptotic expansion of the form*

$$(E_h^{\alpha}t^{\beta-1})(1)=e_0+e_1h+\cdots+e_{N-1}h^{N-1}+O(h^N)+O(h^{\beta}) \tag{3.11}$$

where the coefficients $e_j=e_j(\alpha,\beta,c_0,\cdots,c_j)$ depend analytically on α, β and the coefficients $c_0,\cdots,c_j$ of (3.7).

Proof. a) We need the following auxiliary result: The convolution of two sequences $u_n = O(n^\mu)$ and $v_n = O(n^\nu)$ with $\nu < \min\{-1, \mu-1\}$ satisfies

$$\sum_{j=0}^{n} u_{n-j} v_j = O(n^\mu). \tag{3.12}$$

This is seen from

$$\left| \sum_{j=0}^{n} u_{n-j} v_j \right| \leqq |u_n v_0| + |u_0 v_n| + M n^\mu \sum_{j=1}^{n-1} \left(1 - \frac{j}{n}\right)^\mu j^\nu$$

and

$$\left(1 - \frac{j}{n}\right)^\mu \leqq \begin{cases} 1 & \text{if } \mu \geqq 0 \\ (j+1)^{-\mu} & \text{if } \mu < 0 \end{cases} \quad \text{for } 1 \leqq j \leqq n-1.$$

b) For $\beta \neq 0, -1, -2, \cdots$ we obtain from (3.1) an asymptotic expansion

$$n^{\beta-1} = b_0 (-1)^n \binom{-\beta}{n} + b_1 (-1)^n \binom{-\beta+1}{n} + \cdots + b_{N-1} (-1)^n \binom{-\beta+N-1}{n} + O(n^{\beta-1-N}). \tag{3.13}$$

So we have

$$b(\zeta) := \sum_{n=1}^{\infty} n^{\beta-1} \zeta^n = b_0 (1-\zeta)^{-\beta} + b_1 (1-\zeta)^{-\beta+1} + \cdots + b_{N-1} (1-\zeta)^{-\beta+N-1} + s(\zeta)$$

where the coefficients s_n of $s(\zeta)$ satisfy

$$s_n = O(n^{\beta-1-N}). \tag{3.14}$$

c) We have to study the expression

$$h^\alpha \sum_{j=1}^{n} \omega_{n-j} (jh)^{\beta-1} \qquad (hn = 1).$$

$y_n = \sum_{j=1}^{n} \omega_{n-j} j^{\beta-1}$ is the nth coefficient of $y(\zeta) = \omega(\zeta) b(\zeta)$ which by inserting (3.7) and (3.13) can be written as

$$y(\zeta) = d_0 (1-\zeta)^{-(\alpha+\beta)} + d_1 (1-\zeta)^{-(\alpha+\beta)+1} + \cdots + d_{2N-2} (1-\zeta)^{-(\alpha+\beta)+2N-2} + \omega(\zeta) s(\zeta) + [b(\zeta) - s(\zeta)](1-\zeta)^N r(\zeta)$$

where $d_k = \sum_{j=0}^{k} b_{k-j} c_j$.

If N is chosen sufficiently large, then the coefficients of $\omega(\zeta) s(\zeta)$ and

$$[b(\zeta) - s(\zeta)](1-\zeta)^N r(\zeta) = \left[b_0 (1-\zeta)^{-\beta+N} + \cdots + b_{N-1} (1-\zeta)^{-\beta+2N-1} \right] r(\zeta)$$

are $O(n^{\alpha-1})$ by (3.1), (3.8), (3.14) and (3.12).

By (3.1) we have therefore

$$y_n = \tilde{e}_0 n^{\alpha+\beta-1} + e_1 n^{(\alpha+\beta-1)-1} + \cdots + e_N n^{(\alpha+\beta-1)-N} + O(n^{\alpha-1}).$$

This gives the desired result for

$$\left(E_h^\alpha t^{\beta-1}\right)(1)=h^{\alpha+\beta-1}y_n-\left(I^\alpha t^{\beta-1}\right)(1) \qquad (hn=1). \qquad \square$$

In Lemma 3.8 below we shall show that $e_0=\cdots=e_{p-1}=0$ if the method is stable and consistent of order p. First, we need two auxiliary results in which we restrict our attention to $\operatorname{Re}\alpha>0$.

LEMMA 3.6. *Let* $\operatorname{Re}\alpha>0$.

If $\left(E_h^\alpha t^{p-1}\right)(1)=O(h^p)$, *then* $\left(E_h^\alpha t^{\beta-1}\right)(1)=O(h^p)$ *for all* $\operatorname{Re}\beta>p$.

Proof. Let $\beta=p+\mu$. By (1.4),

$$t^{\beta-1}=\frac{\Gamma(p+\mu)}{\Gamma(p)\Gamma(\mu)}t^{p-1}*t^{\mu-1}.$$

By (3.3),

$$\left(E_h^\alpha t^{p-1}\right)(x)=O(x^{\alpha-1}h^p).$$

By (3.2),

$$E_h^\alpha\left(t^{p-1}*t^{\mu-1}\right)(1)=\left(E_h^\alpha t^{p-1}*t^{\mu-1}\right)(1)=O(h^p).$$

Hence also

$$\left(E_h^\alpha t^{\beta-1}\right)(1)=O(h^p). \qquad \square$$

Remark. $E_h^\alpha t^{p-1}$ is the Peano kernel of the quadrature ω.

LEMMA 3.7. *Let* $\operatorname{Re}\alpha>0$. *There exist numbers* $\tilde\gamma_0,\tilde\gamma_1,\tilde\gamma_2,\cdots$ (*independent of* ω) *such that the following equivalence holds for stable* ω:

$$\left(E_h^\alpha t^{q-1}\right)(1)=O(h^q) \quad \text{for } q=1,2,\cdots,p \tag{3.15}$$

if and only if the coefficients c_i *of* (3.7) *satisfy*

$$c_i=\tilde\gamma_i \quad \text{for } i=0,1,\cdots,p-1. \tag{3.16}$$

Proof. The proof proceeds by induction on p. Trivially the statement holds for $p=0$.

Assume now that Lemma 3.7 has already been proved up to order p. We shall prove it for $p+1$.

Let either of (3.15) or (3.16) hold. By the induction hypothesis, it suffices to show that c_p can be uniquely chosen such that

$$\left(E_h^\alpha t^p\right)(1)=O(h^{p+1}).$$

From Lemma 3.6 (and from Lemma 3.5 for $p=0$) we know already

$$\left(E_h^\alpha t^p\right)(1)=O(h^p). \tag{3.17}$$

For any integer n we may write

$$n^p=\sum_{k=1}^{p+1}b_k\binom{n+k-1}{n}=\sum_{k=1}^{p+1}b_k(-1)^n\binom{-k}{n}$$

so that (with $hn=1$)

$$\left(\Omega_h^\alpha t^p\right)(1)=h^\alpha\sum_{j=0}^{n}\omega_j(n-j)^ph^p=h^{p+\alpha}\sum_{k=1}^{p+1}b_k\sum_{j=0}^{n}\omega_j(-1)^{n-j}\binom{-k}{n-j}.$$

The inner sum is the nth coefficient of

$$\frac{\omega(\zeta)}{(1-\zeta)^k}=\tilde{\gamma}_0(1-\zeta)^{-\alpha-k}+\cdots+\tilde{\gamma}_{p-1}(1-\zeta)^{-\alpha+p-1-k}$$
$$+c_p(1-\zeta)^{-\alpha+p-k}+(1-\zeta)^{p+1-k}r(\zeta).$$

Using (3.1), (3.8) and (3.17) we obtain

$$(E_h^\alpha t^p)(1)=\frac{c_p-\tilde{\gamma}_p}{\Gamma(\alpha+1)}h^p+O(h^{p+1})$$

where $\tilde{\gamma}_p$ depends only on α and $\tilde{\gamma}_0,\cdots,\tilde{\gamma}_{p-1}$. Hence (3.15) holds for $p+1$ instead of p if and only if additionally $c_p=\tilde{\gamma}_p$. □

We have now arrived at the final step of the proof.

LEMMA 3.8. *Let $\alpha\in\mathbb{C}$. If ω is stable and consistent of order p, then it is also convergent of order p.*

Proof. Let first $\mathrm{Re}\,\alpha>0$. Since (3.15) implies consistency of order p by Lemma 3.1, the numbers $\tilde{\gamma}_i$ of Lemma 3.7 and $\gamma_i=\gamma_i(\alpha)$ of Lemma 3.2 are identical.

By Lemmas 3.5 and 3.6 we have then for $\mathrm{Re}\,\alpha>0$, $\mathrm{Re}\,\beta>p$

$$e_j\big(\alpha,\beta,\gamma_0(\alpha),\cdots,\gamma_j(\alpha)\big)=0\qquad(j=0,\cdots,p-1).$$

By analyticity, this holds then for *all* α, β.

If the method is consistent of order p, we have by Lemma 3.2 $c_i=\gamma_i(\alpha)$ for $i=0,\cdots,p-1$. Now Lemma 3.5 gives the result. □

We have thus completed the proof of Theorem 2.5. □

Proof of Theorem 2.6. The linear multistep method is consistent of order p, i.e.

$$h\omega(e^{-h})=1+O(h^p).$$

Taking this relation to power α yields

$$h^\alpha\omega^\alpha(e^{-h})=1+O(h^p),$$

so that ω^α is consistent of order p for I^α. Under the given assumptions on (ρ,σ) we can write

$$\omega(\zeta)=\frac{\sigma(\zeta^{-1})}{\rho(\zeta^{-1})}=\prod_{i=0}^{r}(1-\zeta_i\zeta)^{-1}v(\zeta)$$

where $v(\zeta)$ is analytic and without zeros in a neighbourhood of $|\zeta|\leqq 1$, and ζ_i are the zeros of $\rho(\zeta)$ on the unit circle. Hence

$$\omega^\alpha(\zeta)=\prod_{i=0}^{r}(1-\zeta_i\zeta)^{-\alpha}u(\zeta)$$

where $u(\zeta)=v(\zeta)^\alpha$ is analytic in a neighbourhood of $|\zeta|\leqq 1$. By (3.9) and (3.10),

$$\omega_n^{(\alpha)}=O(n^{\alpha-1})$$

so that ω^α is stable. Now Theorem 2.5 completes the proof. □

Proof of Theorem 2.4. The proof is based on a Peano kernel technique similar as in Lemmas 3.1 and 3.6.

(i) Fix $\beta \neq 0, -1, -2, \cdots$. Let the integer m such that

$$\mathrm{Re}(m+\beta-1) \leqq p < \mathrm{Re}(m+\beta).$$

A suitable starting quadrature can be chosen by putting

$$(3.18) \quad h^\alpha \sum_{j=1}^{m} w_{nj} (jh)^{q+\beta-1} + \left(E_h^\alpha t^{q+\beta-1}\right)(1) = 0 \qquad (q=0,1,\cdots,m-1;\ hn=1).$$

This gives a Vandermonde type system of equations for w_{nj} $(j=1,\cdots,m)$,

$$\sum_{j=1}^{m} w_{nj} j^{q+\beta-1} = O(n^{\alpha-1}) \quad \text{by (2.4)}.$$

Hence also

$$(3.19) \qquad w_{nj} = O(n^{\alpha-1})$$

as desired.

Let $f(x) = x^{\beta-1} g(x)$, g sufficiently differentiable. Expanding f as a fractional Taylor series with Bernoulli remainder term gives (let $f^{(\mu)} = I^{-\mu} f$)

$$(3.20) \quad f(x) = \sum_{q=0}^{N} \frac{f^{(q+\beta-1)}(0)}{\Gamma(q+\beta)} x^{q+\beta-1} + \frac{1}{\Gamma(N+\beta)} \left(t^{N+\beta-1} * f^{(N+\beta)}\right)(x).$$

If $\mathrm{Re}(N+\beta-1) > p$, then (3.3) and (2.4) yield

$$\left(E_h^\alpha t^{N+\beta-1}\right)(x) = O(x^{N-p+\alpha+\beta-1} h^p).$$

If additionally $\mathrm{Re}(N-p+\alpha+\beta) > 0$, then (3.2) and the boundedness of $f^{(N+\beta)}$ give

$$(3.21) \qquad \begin{aligned} E_h^\alpha \left(t^{N+\beta-1} * f^{(N+\beta)}\right)(x) &= \left(E_h^\alpha t^{N+\beta-1} * f^{(N+\beta)}\right)(x) \\ &= O(x^{N-p+\alpha+\beta} h^p) \end{aligned}$$

for bounded x.

By our choice of the starting quadrature ((3.18), (3.19)), by the homogeneity relation (3.3) and by (3.20), (3.21) we have

$$\begin{aligned} I_h^\alpha f(x) - I^\alpha f(x) &= E_h^\alpha f(x) + h^\alpha \sum_{j=1}^{m} w_{nj} f(jh) \\ &= O(x^{m-p+\alpha+\beta-1} h^p) + \cdots + O(x^{N-p+\alpha+\beta-1} h^p) + O(x^{N-p+\alpha+\beta} h^p) \\ &= O(x^{m-p+\alpha+\beta-1} h^p) \quad \text{uniformly for bounded } x. \end{aligned}$$

This gives (i) of Theorem 2.4 (note $\tilde{\beta} = m-p+\beta$).

(ii) If m is (3.18) is replaced by $l > m$ with $\mathrm{Re}(l-p+\alpha+\beta-1) \geqq 0$, then the corresponding starting quadrature weights satisfy

$$w_{nj} = O(n^{l-1-p+\alpha+\beta-1}),$$

and the same arguments as above show

$$I_h^\alpha f(x) - I^\alpha f(x) = O(h^p) \quad \text{uniformly for bounded } x. \qquad \square$$

4. Implementation.

4.1. Weights of fractional linear multistep methods ω^α. The coefficients of ω^α, defined by (2.10), are computed most efficiently by Fast Fourier Transform (FFT) techniques for formal power series as described by Henrici [9, §5]. The weights $\omega_0^{(\alpha)}, \cdots, \omega_{N-1}^{(\alpha)}$ are thus obtained using only $O(N \log N)$ additions and multiplications.

4.2. Starting quadrature weights w_{nj}. Multiplying (3.18) by $n^{q+\alpha+\beta-1}$ and using (1.4) we obtain

(4.1)
$$\sum_{j=1}^{s} w_{nj} j^{q+\beta-1} = \frac{\Gamma(q+\beta)}{\Gamma(\alpha+q+\beta)} n^{q+\alpha+\beta-1} - \sum_{j=1}^{n} \omega_{n-j} j^{q+\beta-1} \qquad (q=0,\cdots,s-1).$$

Exploiting the convolution structure of the right-hand side, the weights $w_{nj} (n = 1,\cdots,N$ and $j=1,\cdots,s)$ can be computed from (4.1) with $O(N \log N)$ operations, using FFT-techniques (cf. [9]).

The starting quadrature of (ii) in Theorem 2.4 can be used on short intervals. If w_{nj} of (ii) do not satisfy (2.5), then they dominate ω_n for large n, and errors in the evaluation of $f(jh)$ $(j=1,\cdots,s)$ are unduly magnified.

(As a marginal note: For large n the right-hand side of (4.1) can be computed only with large relative error, due to cancellation of leading digits. Moreover the Vandermonde system is ill-conditioned for large s. Hence the weights w_{nj} are computed with possibly low accuracy. This does, however, *not* affect the accuracy of the quadrature, since it is only important that (4.1) holds up to machine precision).

4.3. Computation of $\Omega_h^\alpha f$. After $f_j = f(jh)$ have been evaluated, the values of the convolution $\Omega_h^\alpha f(nh) = h^\alpha \sum_{j=1}^{n} \omega_{n-j} f_j (n=1,\cdots,N)$ can be computed simultaneously by FFT-techniques with only $O(N \log N)$ operations.

5. Applications and numerical examples.

5.1. Abel's integral equation. Historically, the first application of fractional calculus was probably given by Abel in his study of the tautochrone problem ([1], see also [18, p. 183]). This led him to the integral equation

$$\frac{1}{\sqrt{\pi}} \int_0^x (x-s)^{-1/2} y(s)\, ds = f(x),$$

the solution of which he found to be

$$y(x) = I^{-1/2} f(x).$$

For our numerical experiments we have used the $(\mathrm{BDF3})^{-1/2}$ method (third order backward differentiation formula to power $-\frac{1}{2}$, see Example 2.8). We give the results for the function

$$f(x) = \frac{x}{1+x}.$$

The exact solution is then given by (see [18, p. 121])

(5.1)
$$I^{-1/2} \frac{x}{1+x} = \frac{2}{\sqrt{\pi}} \sqrt{x}\, {}_2F_1\left(1, 2; \frac{3}{2}; -x\right),$$

where ${}_2F_1$ denotes the hypergeometric function. The solution at $x=1$ is $y(1)=0.4579033863$. The numerical results are given in Table 2.

TABLE 2

h	numerical solution	error	error/h^3
0.04	0.4579085018	$0.512_{10^{-5}}$	0.0799
0.02	0.4579040377	$0.651_{10^{-6}}$	0.0814
0.01	0.4579034683	$0.820_{10^{-7}}$	0.0820

5.2. Diffusion problems. As a simple example, consider the heat equation in a half-space

$$u_t = u_{xx} \qquad (x>0,\ t>0)$$

with initial condition

$$u(x,0)=0 \qquad (x>0),$$

with boundary conditions

$$u(\infty,t)=0 \qquad (t>0)$$

and either

(i) $u(0,t)=f(t)$ $(t>0)$ or

(ii) $u_x(0,t)=g(t)$ $(t>0)$ or

(iii) $u_x(0,t)=G(u(0,t))$ $(t>0)$.

The solution at the surface $x=0$ satisfies (cf. [2, App. 2 to Ch. V])

$$u(0,t) = -\frac{1}{\sqrt{\pi}}\int_0^t (t-s)^{-1/2} u_x(0,s)\,ds \qquad (t>0).$$

For boundary conditions (ii) the surface temperature $u(0,t)$ is thus obtained as $-I^{1/2}g(t)$. For boundary conditions (i) this formula is a first kind Abel integral equation for the surface flux $u_x(0,t)$, which hence equals $-I^{-1/2}f(t)$. In case (iii) we obtain a second kind Abel integral equation for $u(0,t)$. The application of fractional linear multistep methods to such equations is discussed in the author's paper [16]. The solution $u(x,t)$ can be recovered from the surface flux by

$$u(x,t) = -\frac{1}{\sqrt{\pi}}\int_0^t (t-s)^{-1/2} \exp\left(\frac{-x^2}{4(t-s)}\right) u_x(0,s)\,ds.$$

As a numerical example related to (ii), we have used the $(\mathrm{BDF4})^{1/2}$ method (see Example 2.8) to compute

$$I^{1/2}\frac{\sin\sqrt{t}}{\sqrt{\pi}} = \sqrt{t}\,J_1(\sqrt{t}), \tag{5.2}$$

where J_1 denotes the Bessel function (see [18, p. 124]). At $t=1$ the solution is $J_1(1)=0.4400505857449$. The numerical results are given in Table 3.

TABLE 3

h	numerical solution	error	error/h^4
0.04	0.4400505854008	$-0.344_{10^{-9}}$	$-0.134_{10^{-3}}$
0.02	0.4400505857240	$-0.209_{10^{-10}}$	$-0.130_{10^{-3}}$
0.01	0.4400505857436	$-0.128_{10^{-11}}$	$-0.127_{10^{-3}}$

5.3. Special functions. The relations (5.1) and (5.2) are special cases of

$$ {}_2F_1(a,b;c;x)=\frac{\Gamma(c)x^{1-c}}{\Gamma(b)}I^{c-b}\left[x^{b-1}(1-x)^{-a}\right] \quad \text{and}$$

$$ J_\mu(\sqrt{x})=\frac{2}{\sqrt{\pi}}(2\sqrt{x})^{-\mu}I^{\mu-1/2}\sin\sqrt{x}\,.$$

Among the special functions which can be represented as fractional integrals of simpler functions are: hypergeometric functions, confluent and generalized hypergeometric functions, Bessel and Struve functions, Legendre functions, elliptic integrals etc. (see [12], [18]). Convolution quadratures for their computation are particularly effective if one is interested in obtaining many values on a grid simultaneously.

Acknowledgments. The author wishes to thank E. Hairer and G. Gienger for helpful discussions, and G. Bader, P. Deuflhard and U. Nowak for their interest in this work.

REFERENCES

[1] N. H. ABEL, *Solution de quelques problèmes à l'aide d'intégrales définites* (1823), in Oeuvres complètes, Vol. 1, Grondahl, Christiania, Norway, 1881, pp. 16–18.

[2] R. COURANT AND D. HILBERT, *Methods of Mathematical Physics, Vol.* II, Wiley-Interscience, New York, London, 1962.

[3] G. DAHLQUIST, *Convergence and stability in the numerical integration of ordinary differential equations*, Math. Scand. 4 (1956), pp. 33–53.

[4] _____, *Stability and error bounds in the numerical integration of ordinary differential equations*, Trans. Royal Institute of Technology Stockholm, Nr. 130, 1959.

[5] _____, *On accuracy and unconditional stability of linear multistep methods for second order differential equations*, BIT, 18 (1978), pp. 133–136.

[6] A. ERDÉLYI, ed., *Higher Transcendental Functions* I, McGraw-Hill, New York, Toronto, London, 1953.

[7] A. K. GRÜNWALD, *Über "begrenzte" Derivationen und deren Anwendung*, Z. Math. Phys., 12 (1867), pp. 441–480.

[8] P. HENRICI, *Discrete Variable Methods in Ordinary Differential Equations*, John Wiley, New York, 1962.

[9] _____, *Fast Fourier methods in computational complex analysis*, SIAM Rev., 21 (1979), pp. 481–527.

[10] R. JELTSCH, *Stability on the imaginary axis and A-stability of linear multistep methods*, BIT, 18 (1978), pp. 170–174.

[11] Y. KATZNELSON, *An Introduction to Harmonic Analysis*, John Wiley, New York, 1968.

[12] J. L. LAVOIE, T. J. OSLER AND R. TREMBLAY, *Fractional derivatives and special functions*, SIAM Rev., 18 (1976), pp. 240–268.

[13] A. V. LETNIKOV, *Theory of differentiation of fractional order*, Mat. Sb., 3 (1868), pp. 1–68.

[14] J. LIOUVILLE, *Mémoire sur le calcul des différentielles à indices quelconques*, J. de l'Ecole Polytechnique, 13 (1832), pp. 71–162.

[15] CH. LUBICH, *On the stability of linear multistep methods for Volterra convolution equations*, IMA J. Numer. Anal., 3 (1983), pp. 439–465.

[16] _____, *Fractional linear multistep methods for Abel–Volterra integral equations of the second kind*, Math. Comput., (1985), to appear.

[17] J. MATTHYS, *A-stable linear multistep methods for Volterra integro-differential equations*, Numer. Math., 27 (1976), pp. 85–94.

[18] K. B. OLDHAM AND J. SPANIER, *The Fractional Calculus*, Academic Press, New York, London, 1974.

[19] M. RIESZ, *L'intégrale de Riemann-Liouville et le problème de Cauchy*, Acta Math., 81 (1949), pp. 1–223.

[20] P. H. M. WOLKENFELT, *The numerical analysis of reducible quadrature methods for Volterra integral and integro-differential equations*, thesis, Math. Centrum, Amsterdam, 1981.

SIAM J. MATH. ANAL.
Vol. 17, No. 3, May 1986

013

THE REPRESENTATION OF FUNCTIONS AS LAPLACE AND LAPLACE–STIELTJES TRANSFORMS*

F. J. WILSON†

Abstract. We develop necessary and sufficient conditions for a function to be represented as a Laplace or Laplace–Stieltjes transform by considering the behaviour of the function on a single vertical line. Various kernels, based on ideal inversion kernels for the Fourier transform, are considered and three new inversion formulae for the Laplace transform are developed.

AMS(MOS) subject classification. Primary 44A10

Key words. Laplace transform, inversion kernel

1. Introduction. In this paper we examine criteria for a function $f(w)$ to be represented as a Laplace or Laplace–Stieltjes transform by considering the behaviour of the function on a single vertical line. We consider integral transforms of the form

$$F(t,c,\lambda)=\frac{1}{2\pi}\int_{-\infty}^{\infty} k(t,v,c,\lambda)f(c+iv)\,dv$$

where $k(t,v,c,\lambda)$ is a kernel similar in form to the ideal inversion kernels for the Fourier transform which were examined in [1] and [6].

In [2] Cooper gives necessary and sufficient conditions for a function to be represented as a Laplace transform by considering integral transforms of the form

$$F(t,c,\lambda)=\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{(c+iv)t}l(v,\lambda)f(c+iv)\,dv.$$

We extend these results by considering a wider class of kernels some of which, for example the Post–Widder kernel

$$k(t,v,c,\lambda)=\left(\frac{\lambda}{\lambda-t(c+iv)}\right)^{\lambda+1},$$

cannot be written in the form $e^{(c+iv)t}l(v,\lambda)$. We also prove our results for p in the range $1\leqq p\leqq\infty$ whereas Cooper in [2], because the proofs of his main theorems depend on Fourier transform theory, has to restrict p to the range $1\leqq p\leqq 2$.

We show that, under certain conditions on $f(w)$ and the kernels $k(t,v,c,\lambda)$, the boundedness of the set $\{F(t,c,\lambda)e^{-at}\}$ in $L_p(0,\infty)$, $1<p\leqq\infty$, is necessary and sufficient for $f(w)$ to be represented as a Laplace transform and if f is the Laplace transform of a function F then F can be found by considering the weak limit of $F(t,c,\lambda)e^{-at}$ in $L_p(0,\infty)$ as $\lambda\to\infty$. Similarly for $p=1$ we show that the boundedness of the set $\{F(t,c,\lambda)e^{-at}\}$ in $L(0,\infty)$ is necessary and sufficient for $f(w)$ to be represented as a Laplace–Stieltjes transform and if f is the Laplace–Stieltjes transform of a function H then H can be found by considering the limit of $\int_0^t F(u,c,\lambda)\,du$ as $\lambda\to\infty$.

* Received by the editors May 9, 1984.

† Department of Mathematics, University of Wales Institute of Science and Technology, Cardiff CF1 3EU, United Kingdom.

We show that the Weierstrass and Abel kernels, the Cesaro kernel of positive order and the kernel of ordinary convergence satisfy the conditions of the theorems. We then produce three new inversion formulae for the Laplace transform by showing that the following kernels

$$\left(\frac{\lambda}{\lambda - t(c+iv)}\right)^{\lambda+1} \quad \text{the Post–Widder kernel,}$$

$$\left(\frac{\lambda+\theta_\lambda}{\lambda+\theta_\lambda - t(c+iv)}\right)^{\lambda+1} \quad \text{where } \theta_\lambda = o(\lambda) \text{ as } \lambda \to \infty,$$

the extended Post–Widder kernel,

and

$$e^{(c+iv)t}\Gamma\left(1-\frac{iv}{\lambda}\right) \quad \text{the extended Phragmen kernel}$$

also satisfy the conditions of the theorems. Lastly we show that for a particular class of kernels it is possible to determine the abscissa of absolute convergence of the Laplace transform by considering the behaviour of $f(w)$ on any line in the half-plane of absolute convergence.

2. Notation. We write

$$(LF)(w) = \int_0^\infty e^{-wt} F(t)\, dt$$

and

$$(SH)(w) = \int_0^\infty e^{-wt}\, dH(t)$$

for the Laplace and Laplace–Stieltjes transforms.

For any real a we write $F \in L_p[a;(0,\infty)]$ when $F(t)e^{-at} \in L_p(0,\infty)$ and $H \in V[a,(0,\infty)]$ when H is a function of bounded variation over $(0,\infty)$ and

$$\int_0^\infty e^{-at} |dH(t)|$$

is finite. We will denote the dual of $L_p(0,\infty)$ by $L_{p'}(0,\infty)$. For $1 \leqq p < \infty$ we use the following notation for function norms

$$\|\psi(t);(0,\infty)\|_p = \left\{\int_0^\infty |\psi(t)|^p\, dt\right\}^{1/p}$$

and for $p = \infty$

$$\|\psi(t);(0,\infty)\|_\infty = \operatorname*{ess\,sup}_{0<t<\infty} |\psi(t)|$$

We write $e^{-(c+iv)t}k(t,v,c,\lambda) \to 1$ boundedly for almost all (v,t) in $(-\infty,\infty)\times(0,\infty)$ as $\lambda \to \infty$ when

$$\operatorname*{ess\,sup}_{\substack{-\infty<v<\infty \\ 0<t<\infty}} |e^{-(c+iv)t}k(t,v,c,\lambda)| < M$$

for all $\lambda > \lambda_0$, where M is a constant independent of v and t, and

$$e^{-(c+iv)t}k(t,v,c,\lambda) \to 1$$

as $\lambda \to \infty$ for almost all (v,t) in $(-\infty,\infty)\times(0,\infty)$.

We write

$$\int_0^\infty g(t,x,\lambda)\begin{Bmatrix} dt \\ dx \end{Bmatrix} \text{ is bounded uniformly in } \begin{Bmatrix} x \\ t \end{Bmatrix} \text{ and } \lambda$$

when $\int_0^\infty g(t,x,\lambda)\,dt$ is bounded uniformly in x and λ and $\int_0^\infty g(t,x,\lambda)\,dx$ is bounded uniformly in t and λ.

For the Fourier transform of $k(t,v,c,\lambda)$ with respect to v we write

$$K(t,x,c,\lambda) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-ivx}k(t,v,c,\lambda)\,dv.$$

We will denote the characteristic function of the interval (a,b) by $\chi_{(a,b)}(v)$.

In all theorems we are only concerned with $\lambda \to \infty$ so to avoid any difficulties with small values of λ we will restrict λ to the range $\lambda > \lambda_0$.

3. Necessary conditions for the representation of functions as Laplace or Laplace–Stieltjes transforms. We now show that, under certain conditions on the kernel $k(t,v,c,\lambda)$, the boundedness of the integral transform

$$F(t,c,\lambda)e^{-at} = \frac{e^{-at}}{2\pi} \int_{-\infty}^{\infty} k(t,c,v,\lambda)f(c+iv)\,dv$$

in $L_p(0,\infty)$ is a necessary condition for a function $f(w)$ to be represented as a Laplace or Laplace–Stieltjes transform.

For the theorems dealing with necessary conditions there are no restrictions on the relative values of the real numbers a and c apart from the restrictions imposed by the given function $F(t)$ and the kernel $k(t,v,c,\lambda)$ but in §4, where we prove the corresponding sufficiency theorems, we will require that $a \leqq c$.

Firstly for the case $1 < p < \infty$.

THEOREM 1. *Let $f = LF$, where $F \in L_p[a;(0,\infty)]$ and $1 < p < \infty$, and let c be a real number such that the Laplace transform, $f(w)$, converges absolutely for $w = c + iv$. As a function of v let $k(t,v,c,\lambda) \in L(-\infty,\infty)$ for all $t > 0$ and $\lambda > \lambda_0$.*

Let $\int_0^\infty e^{-at}e^{(a-c)x}|K(t,x,c,\lambda)|\begin{Bmatrix} dt \\ dx \end{Bmatrix}$ be bounded uniformly in $\begin{Bmatrix} x \\ t \end{Bmatrix}$ and λ for all $\begin{Bmatrix} x>0 \\ t>0 \end{Bmatrix}$ and $\lambda > \lambda_0$. Then there exists a constant M_p such that

$$\|F(t,c,\lambda)e^{-at};(0,\infty)\|_p \leqq M_p$$

uniformly in λ for all $\lambda > \lambda_0$.

Proof. By hypothesis there exist constants N_1 and N_2 such that

$$\int_0^\infty e^{-at}e^{(a-c)x}|K(t,x,c,\lambda)|\,dt \leqq N_1$$

and

$$\int_0^\infty e^{-at}e^{(a-c)x}|K(t,x,c,\lambda)|\,dx \leqq N_2.$$

Now

$$F(t,c,\lambda)e^{-at} = \frac{e^{-at}}{2\pi}\int_{-\infty}^{\infty} k(t,v,c,\lambda)f(c+iv)\,dv$$

$$= \frac{e^{-at}}{2\pi}\int_{-\infty}^{\infty} k(t,v,c,\lambda)\,dv\int_0^{\infty} e^{-(c+iv)x}F(x)\,dx$$

$$= \frac{1}{\sqrt{2\pi}}\int_0^{\infty} F(x)e^{-ax}e^{-at}e^{(a-c)x}K(t,x,c,\lambda)\,dx.$$

Hence, using Holder's inequality,

$$\|F(t,c,\lambda)e^{-at};(0,\infty)\|_p^p$$

$$\leqq (2\pi)^{-p/2}\int_0^{\infty} dt\left[\left\{\int_0^{\infty}|F(x)e^{-ax}|^p e^{-at}e^{(a-c)x}|K(t,x,c,\lambda)|\,dx\right\}\right.$$

$$\left.\cdot\left\{\int_0^{\infty} e^{-at}e^{(a-c)x}|K(t,x,c,\lambda)|\,dx\right\}^{p/p'}\right]$$

$$\leqq (2\pi)^{-p/2}N_2^{p/p'}\int_0^{\infty} dt\int_0^{\infty}|F(x)e^{-ax}|^p e^{-at}e^{(a-c)x}|K(t,x,c,\lambda)|\,dx$$

$$\leqq (2\pi)^{-p/2}N_2^{p/p'}\int_0^{\infty}|F(x)e^{-ax}|^p\,dx\int_0^{\infty} e^{-at}e^{(a-c)x}|K(t,x,c,\lambda)|\,dt$$

$$\leqq (2\pi)^{-p/2}N_2^{p/p'}N_1\int_0^{\infty}|F(x)e^{-ax}|^p\,dx$$

which is bounded, since $F\in L_p[a;(0,\infty)]$, where the two interchanges of integration are justified by Fubini's theorem.

Theorem 1 also holds for $p=\infty$ and $p=1$ but we show that it is possible to prove the necessity theorem under less restrictive conditions on $k(t,v,c,\lambda)$ for these two particular cases.

Secondly for the case $p=\infty$.

THEOREM 2. *Let $f=LF$, where $F\in L_\infty[a;(0,\infty)]$, and let c be a real number such that the Laplace transform, $f(w)$, converges absolutely for $w=c+iv$. As a function of v let $k(t,v,c,\lambda)\in L(-\infty,\infty)$ for all $t>0$ and $\lambda>\lambda_0$. Let*

$$\int_0^{\infty} e^{-at}e^{(a-c)x}|K(t,x,c,\lambda)|\,dx$$

be bounded uniformly in t and λ for all $t>0$ and $\lambda>\lambda_0$. Then there exists a constant M_∞ such that

$$\|F(t,c,\lambda)e^{-at};(0,\infty)\|_\infty \leqq M_\infty$$

uniformly in λ for all $\lambda>\lambda_0$.

Proof. Using the same method as in Theorem 1, we have

$$\begin{aligned}&\|F(t,c,\lambda)e^{-at};(0,\infty)\|_\infty\\&\quad=\frac{1}{\sqrt{2\pi}}\operatorname*{ess\,sup}_{0<t<\infty}\left|\int_0^\infty F(x)e^{-ax}e^{-at}e^{(a-c)x}K(t,x,c,\lambda)\,dx\right|\\&\quad\leqq\frac{1}{\sqrt{2\pi}}\|F(x)e^{-ax};(0,\infty)\|_\infty\operatorname*{ess\,sup}_{0<t<\infty}\int_0^\infty e^{-at}e^{(a-c)x}|K(t,x,c,\lambda)|\,dx\end{aligned}$$

which is bounded uniformly in λ.

Thirdly for the case $p=1$ we need to consider Laplace–Stieltjes transforms.

THEOREM 3. *Let $f=SH$, where $H\in V[a;(0,\infty)]$, and let c be a real number such that the Laplace–Stieltjes transform, $f(w)$, converges absolutely for $w=c+iv$. As a function of v let $k(t,v,c,\lambda)\in L(-\infty,\infty)$ for all $t>0$ and $\lambda>\lambda_0$. Let*

$$\int_0^\infty e^{-at}e^{(a-c)x}|K(t,x,c,\lambda)|\,dt$$

be bounded uniformly in x and λ for all $x>0$ and $\lambda>\lambda_0$. Then there exists a constant M_1 such that

$$\|F(t,c,\lambda)e^{-at};(0,\infty)\|_1\leqq M_1$$

uniformly in λ for all $\lambda>\lambda_0$.

Proof. Using the same method as in Theorem 1, we have

$$\begin{aligned}&\|F(t,c,\lambda)e^{-at};(0,\infty)\|_1\\&\qquad=\frac{1}{\sqrt{2\pi}}\int_0^\infty dt\left|\int_0^\infty e^{-at}e^{(a-c)x}K(t,x,c,\lambda)e^{-ax}\,dH(x)\right|\\&\qquad\leqq\frac{1}{\sqrt{2\pi}}\int_0^\infty e^{-ax}|dH(x)|\int_0^\infty e^{-at}e^{(a-c)x}|K(t,x,c,\lambda)|\,dt\end{aligned}$$

which is bounded uniformly in λ. The interchange in the order of integration being justified by Fubini's theorem.

Whilst these three theorems are adequate for most of the particular kernels that we will be considering, the kernel of ordinary convergence, $k(t,v,c,\lambda)=e^{(c+iv)t}\chi_{(-\lambda,\lambda)}(v)$, does not satisfy the conditions of the theorems, since, even for the case $a=c$,

$$\int_0^\infty e^{-at}e^{(a-c)x}|K(t,x,c,\lambda)|\,dx$$

is not bounded.

For the kernel of ordinary convergence we need the following necessity theorem.

THEOREM 4. *Let $f=LF$, where $F\in L_p[a;(0,\infty)]$ and $p>1$, and let c be a real number such that the Laplace transform, $f(w)$, converges absolutely for $w=c+iv$. As a function of v let $k(t,v,c,\lambda)\in L(-\infty,\infty)$ for all $t>0$ and $\lambda>\lambda_0$.*

Let the transformation S_λ be defined by

$$(S_\lambda g)(t)=\frac{1}{\sqrt{2\pi}}\int_0^\infty g(x)e^{-at}e^{(a-c)x}K(t,x,c,\lambda)\,dx$$

where $g\in L_p(0,\infty)$.

Then, if the set of transforms $\{S_\lambda\}$ *forms a bounded set of transforms from* $L_p(0,\infty)$ *into itself, there exists a constant,* M_p, *such that*

$$\|F(t,c,\lambda)e^{-at};(0,\infty)\|_p \leqq M_p$$

uniformly in λ *for all* $\lambda > \lambda_0$.

Proof. The proof follows that of Theorems 1 and 2 up to the equation

$$F(t,c,\lambda)e^{-at} = \frac{1}{\sqrt{2\pi}} \int_0^\infty F(x)e^{-ax}e^{-at}e^{(a-c)x}K(t,x,c,\lambda)\,dx$$

and the result then follows directly from the hypothesis.

A similar theorem can be proved for the case $p=1$ but we do not need this theorem as the kernel of ordinary convergence does not satisfy the conditions of Theorem 4 when $p=1$ [1, p. 291].

4. Sufficient conditions for the representation of functions as Laplace or Laplace–Stieltjes transforms. We now show that, under certain conditions on $f(w)$ and the kernel $k(t,v,c,\lambda)$, the boundedness of the integral transform

$$F(t,c,\lambda)e^{-at} = \frac{e^{-at}}{2\pi} \int_{-\infty}^\infty k(t,v,c,\lambda)f(c+iv)\,dv$$

in $L_p(0,\infty)$ is a sufficient condition for a function $f(w)$ to be represented as a Laplace or Laplace–Stieltjes transform. As already mentioned in §3 we require that the value of a be restricted to $a \leqq c$.

Firstly we prove the theorem which gives sufficient conditions for a function to be represented as a Laplace transform for p in the range $1 < p \leqq \infty$.

THEOREM 5. *Let* $f(w)$ *be holomorphic in* $\mathrm{Re}\,w > c$ *and continuous in* $\mathrm{Re}\,w \geqq c$. *For every* $\delta > 0$ *let there be an* $A(\delta)$ *such that*

$$|f(w)| < Ae^{\delta|w-c|^2}$$

throughout the half-plane $\mathrm{Re}\,w > c$. *For some m let*

$$f(c+iv) = O(v^m)$$

as $|v| \to \infty$, *for every* $\delta > 0$ *let*

$$f(u) = O(e^{\delta u})$$

as $u \to \infty$ *and let*

$$|f(c+re^{i\theta})| \to 0$$

as $r \to \infty$ *for some* θ *in the range* $-\pi/2 < \theta < \pi/2$.

Let $f(c+iv) \in L(-\infty,\infty)$ *and let* $e^{-(c+iv)t}k(t,v,c,\lambda) \to 1$ *boundedly for almost all* (v,t) *in* $(-\infty,\infty)\times(0,\infty)$ *as* $\lambda \to \infty$.

For some $a \leqq c$ *let*

$$\|F(t,c,\lambda)e^{-at};(0,\infty)\|_p \leqq M_p$$

uniformly in λ *for all* $\lambda > \lambda_0$.

Then, for $1 < p \leqq \infty$, $f = LF$ *where* $F \in L_p[a;(0,\infty)]$ *and* $F(t)e^{-at}$ *is the weak limit in* $L_p(0,\infty)$ *of* $F(t,c,\lambda)e^{-at}$ *as* $\lambda \to \infty$.

Proof. Let n be a nonnegative integer such that $n > m$ then for $\operatorname{Re} w > c$

$$\int_0^\infty t^n e^{-wt} F(t,c,\lambda)\,dt = \frac{1}{2\pi}\int_0^\infty t^n e^{-wt}\,dt \int_{-\infty}^\infty k(t,v,c,\lambda) f(c+iv)\,dv$$

$$\to \frac{1}{2\pi}\int_{-\infty}^\infty f(c+iv)\,dv \int_0^\infty t^n e^{-wt} e^{(c+iv)t}\,dt$$

as $\lambda \to \infty$, where the interchange in the order of integration is justified by Fubini's theorem and the limiting process by dominated convergence. Therefore

$$\int_0^\infty t^n e^{-wt} F(t,c,\lambda)\,dt \to \frac{n!}{2\pi}\int_{-\infty}^\infty \frac{f(c+iv)}{[w-(c+iv)]^{n+1}}\,dv \tag{4.1}$$

as $\lambda \to \infty$ for $\operatorname{Re} w > c$. Now by hypothesis

$$\| F(t,c,\lambda) e^{-at}; (0,\infty) \|_p \leqq M_p$$

uniformly in λ. This is a bounded set in the dual of the Banach space $L_{p'}(0,\infty)$ and is therefore relatively weakly compact. Hence there exists $F(t,c)$ such that $F(t,c)e^{-at} \in L_p(0,\infty)$ and $F(t,c)e^{-at}$ is a weak limiting point of $F(t,c,\lambda)e^{-at}$ in $L_p(0,\infty)$ as $\lambda \to \infty$. Since

$$t^n e^{-wt} e^{at} \in L_{p'}(0,\infty)$$

for $\operatorname{Re} w > a$ then as $\lambda \to \infty$ $\int_0^\infty t^n e^{-wt} F(t,c)\,dt$ is a limiting point of $\int_0^\infty t^n e^{-wt} F(t,c,\lambda)\,dt$. Hence by the above statement and (4.1) we have, for $\operatorname{Re} w > c$,

$$\int_0^\infty t^n e^{-wt} F(t,c)\,dt = \frac{n!}{2\pi}\int_{-\infty}^\infty \frac{f(c+iv)}{[w-(c+iv)]^{n+1}}\,dv$$

$$= (-1)^n f^{(n)}(w)$$

where the use of Cauchy's formula is justified by the hypothesis, since $n > m$ and $(w-c+1)^{-m} f(w)$ is bounded [3, p. 1326].

Because of the uniqueness of Laplace transforms $F(t,c)$ must be independent of the choice of c; therefore letting $F(t) = F(t,c)$, we have

$$\int_0^\infty t^n e^{-wt} F(t)\,dt = (-1)^n f^{(n)}(w)$$

for $\operatorname{Re} w > c$, where $F(t)e^{-at}$ is the weak limit in $L_p(0,\infty)$ of $F(t,c,\lambda)e^{-at}$ as $\lambda \to \infty$.

Define

$$\phi(w) = \int_0^\infty e^{-wt} F(t)\,dt;$$

then $\phi(w)$ is holomorphic in $\operatorname{Re} w > c$ and

$$\phi^{(n)}(w) = (-1)^n \int_0^\infty t^n e^{-wt} F(t)\,dt = f^{(n)}(w).$$

Therefore for $\operatorname{Re} w > c$ $\phi(w) = f(w) + p(w)$ where $p(w)$ is a polynomial of degree at most $(n-1)$. Letting $w = c + re^{i\theta}$, we have $|\phi(c+re^{i\theta})| \to 0$ as $r \to \infty$ for $-\pi/2 < \theta < \pi/2$, since $F \in L_p[a;(0,\infty)]$ and, by hypothesis, $|f(c+re^{i\theta})| \to 0$ as $r \to \infty$ for some θ in the range $-\pi/2 < \theta < \pi/2$. Hence for $\operatorname{Re} w > c$ $p(w) = 0$ and therefore $\phi(w) = f(w)$.

Hence $f(w)=\int_0^\infty e^{-wt}F(t)\,dt$ where $F\in L_p[a;(0,\infty)]$ and $F(t)e^{-at}$ is the weak limit in $L_p(0,\infty)$ of $F(t,c,\lambda)e^{-at}$ as $\lambda\to\infty$.

Secondly for the case $p=1$ we prove the theorem which gives sufficient conditions for a function to be represented as a Laplace–Stieltjes transform.

THEOREM 6. *Let $f(w)$ be holomorphic in $\operatorname{Re}w>c$ and continuous in $\operatorname{Re}w\geqq c$. For every $\delta>0$ let there be an $A(\delta)$ such that*

$$|f(w)|<Ae^{\delta|w-c|^2}$$

throughout the half-plane $\operatorname{Re}w>c$. For some m let

$$f(c+iv)=O(v^m)$$

as $|v|\to\infty$, for every $\delta>0$ let

$$f(u)=O(e^{\delta u})$$

as $u\to\infty$ and let

$$f(c+re^{i\theta})=O(1)$$

as $r\to\infty$ for some θ in the range $-\pi/2<\theta<\pi/2$.

Let $f(c+iv)\in L(-\infty,\infty)$ and let $e^{-(c+iv)t}k(t,v,c,\lambda)\to 1$ boundedly for almost all (v,t) in $(-\infty,\infty)\times(0,\infty)$ as $\lambda\to\infty$.

For some $a\leqq c$ let

$$\|F(t,c,\lambda)e^{-at};(0,\infty)\|_1\leqq M_1$$

uniformly in λ for all $\lambda>\lambda_0$.

Then $f=SH$, where $H\in V[a;(0,\infty)]$ and

$$H(t)-H(0)=\lim_{\lambda\to\infty}\int_0^t F(u,c,\lambda)\,du.$$

Proof. Let n be a nonnegative integer such that $n>m$ then, as in Theorem 5, we can show that for $\operatorname{Re}w>c$ as $\lambda\to\infty$

$$\int_0^\infty t^n e^{-wt}F(t,c,\lambda)\,dt\to\frac{n!}{2\pi}\int_{-\infty}^{\infty}\frac{f(c+iv)}{[w-(c+iv)]^{n+1}}\,dv. \tag{4.2}$$

Now by hypothesis

$$\|F(t,c,\lambda)e^{-at};(0,\infty)\|_1\leqq M_1$$

uniformly in λ. Therefore $\{\int_0^t F(u,c,\lambda)e^{-au}\,du\}$ is a bounded set in the dual of the Banach space $C_0(0,\infty)$ and is therefore relatively weakly compact. Hence there exists $h(t,c)\in V(0,\infty)$ such that $h(t,c)$ is a weak limiting point of $\int_0^t F(u,c,\lambda)e^{-au}\,du$ as $\lambda\to\infty$. Therefore

$$\int_0^\infty t^n e^{-wt}e^{at}\,dh(t,c)\text{ is a limiting point of }\int_0^\infty t^n e^{-wt}F(t,c,\lambda)\,dt$$

as $\lambda\to\infty$.

Hence by the above statement, (4.2) and using Cauchy's formula, as in theorem 5, we have for $\operatorname{Re}w>c$

$$\int_0^\infty t^n e^{-wt}e^{at}\,dh(t,c)=\frac{n!}{2\pi}\int_{-\infty}^{\infty}\frac{f(c+iv)}{[w-(c+iv)]^{n+1}}\,dv=(-1)^n f^{(n)}(w).$$

Because of the uniqueness of normalised Laplace–Stieltjes transforms, $h(t,c)$ must be independent of the choice of c; therefore letting $h(t,c)=h(t)$ we have

$$\int_0^\infty t^n e^{-wt} e^{at}\, dh(t) = (-1)^n f^{(n)}(w)$$

for $\operatorname{Re} w > c$.

Define $\phi(w)=\int_0^\infty e^{-wt}e^{at}\,dh(t)$. Then $\phi(w)$ is holomorphic in $\operatorname{Re} w > c$ and

$$\phi^{(n)}(w) = (-1)^n \int_0^\infty t^n e^{-wt} e^{at}\, dh(t) = f^{(n)}(w).$$

Therefore for $\operatorname{Re} w > c$

$$\phi(w) = f(w) + p(w)$$

where $p(w)$ is a polynomial in w of degree at most $(n-1)$. Letting $w = c + re^{i\theta}$, we have

$$f(c+re^{i\theta}) = O(1) \quad \text{as } r \to \infty \quad \text{for } -\pi/2 < \theta < \pi/2,$$

since $h(t) \in V(0,\infty)$, and, by hypothesis,

$$p(c+re^{i\theta}) = O(1) \quad \text{as } r \to \infty$$

for some θ in the range $-\pi/2 < \theta < \pi/2$. Hence for $\operatorname{Re} w > c$

$$p(w) = A,$$

where A is a constant, and therefore

$$f(w) = \int_0^\infty e^{-wt} e^{at}\, dh(t) - A.$$

Now if we define

$$H(t) - H(0) = \int_0^t e^{au}\, dh(u),$$

then

$$f(w) = \int_0^\infty e^{-wt}\, dH(t) - A$$

and the constant A can be absorbed into $H(t)$ by altering the value of $H(t)$ at the origin.

Hence

$$f(w) = \int_0^\infty e^{-wt}\, dH(t)$$

where $H \in V[a;(0,\infty)]$ and

$$H(t) - H(0) = \lim_{\lambda \to \infty} \int_0^t F(u,c,\lambda)\, du.$$

Most of the particular kernels which we will be considering are of the form

$$k(t,v,c,\lambda) = e^{(c+iv)t} l(v,\lambda).$$

For kernels of this form it is possible in Theorems 5 and 6 to replace the condition $f(c+iv)\in L(-\infty,\infty)$ by the condition $l(v,\lambda)f(c+iv)\in L(-\infty,\infty)$ for all $\lambda>\lambda_0$. If, in addition, p is restricted to the range $1\leqq p\leqq 2$, then the conditions $|f(c+re^{i\theta})|\to 0$ and $f(c+re^{i\theta})=O(1)$ as $r\to\infty$ for some θ in the range $-\pi/2<\theta<\pi/2$ in Theorems 5 and 6 respectively can be omitted [2, p. 230].

5. Particular kernels. We now consider the conditions which must be satisfied by the kernels for them to be used in the theorems that we have proved.

L1. $k(t,v,c,\lambda)\in L(-\infty,\infty)$ as a function of v for all $t>0$ and $\lambda>\lambda_0$.

L2. $e^{-(c+iv)t}k(t,v,c,\lambda)\to 1$ boundedly for almost all (v,t) in $(-\infty,\infty)\times(0,\infty)$ as $\lambda\to\infty$.

L3.

$$\int_0^\infty e^{-at}e^{(a-c)x}|K(t,x,c,\lambda)|\begin{Bmatrix}dt\\dx\end{Bmatrix} \text{ is bounded uniformly in } \begin{Bmatrix}x\\t\end{Bmatrix}$$

$$\text{and } \lambda \text{ for all } \begin{Bmatrix}x>0\\t>0\end{Bmatrix} \text{ and } \lambda>\lambda_0.$$

L4. The set of transforms $\{S_\lambda\}$ forms a bounded set of transformations from $L_p(0,\infty)$ into itself, where the transformation S_λ is defined by

$$(S_\lambda g)(t)=\frac{1}{\sqrt{2\pi}}\int_0^\infty g(x)e^{-at}e^{(a-c)x}K(t,x,c,\lambda)\,dx$$

where $g\in L_p(0,\infty)$.

We say that $k(t,v,c,\lambda)$ is an $L(a\leqq c)$ kernel if it satisfies the conditions L1, L2 and either L3 or L4. If the conditions are only satisfied when $a=c$, then we write $k(t,v,c,\lambda)$ is an $L(a=c)$ kernel.

The Weierstrass kernel

$$k(t,v,c,\lambda)=e^{(c+iv)t}e^{-v^2/\lambda^2}$$

satisfies L1 and L2.

$$K(t,x,c,\lambda)=\frac{e^{ct}}{\sqrt{2\pi}}\int_{-\infty}^{\infty}e^{-iv(x-t)}e^{-v^2/\lambda^2}\,dv$$

$$=\frac{\lambda e^{ct}}{\sqrt{2}}\exp\left\{\frac{-\lambda^2(x-t)^2}{4}\right\}.$$

Therefore letting $y=\lambda(x-t)/\sqrt{2}$,

$$\int_0^\infty e^{-at}e^{(a-c)x}|K(t,x,c,\lambda)|\,dt$$

$$=\exp\left\{\frac{(a-c)^2}{\lambda^2}\right\}\int_{-\infty}^{x\lambda/\sqrt{2}}\exp\left\{-\frac{1}{2}\left[y-\sqrt{2}(a-c)/\lambda\right]^2\right\}dy$$

$$\leqq\sqrt{2\pi}\exp\left\{\frac{(a-c)^2}{\lambda^2}\right\}.$$

Similarly,

$$\int_0^\infty e^{-at}e^{(a-c)x}|K(t,x,c,\lambda)|dx$$

$$=\exp\left\{\frac{(a-c)^2}{\lambda^2}\right\}\int_{-t\lambda/\sqrt{2}}^\infty \exp\left\{-\frac{1}{2}\left[y-\sqrt{2}(a-c)/\lambda\right]^2\right\}dy$$

$$\leqq\sqrt{2\pi}\exp\left\{\frac{(a-c)^2}{\lambda^2}\right\}$$

and therefore L3 is satisfied. Hence the Weierstrass kernel is an L $(a\leqq c)$ kernel.

The Abel kernel

$$k(t,v,c,\lambda)=e^{(c+iv)t}e^{-|v|/\lambda}$$

satisfies L1 and L2.

$$K(t,x,c,\lambda)=\frac{c^{ct}}{\sqrt{2\pi}}\int_{-\infty}^\infty e^{-iv(x-t)}e^{-|v|/\lambda}dv$$

$$=\frac{2\lambda e^{ct}}{\sqrt{2\pi}}\frac{1}{1+\lambda^2(x-t)^2}.$$

Therefore letting $y=x-t$,

$$\int_0^\infty e^{-at}e^{(a-c)x}|K(t,x,c,\lambda)|dt=\sqrt{\frac{2}{\pi}}\int_{-\infty}^x e^{(a-c)y}\frac{\lambda}{1+\lambda^2y^2}dy$$

which cannot be bounded if $a<c$. We will therefore only consider the possibility $a=c$. In which case

$$\int_0^\infty e^{-ct}|K(t,x,c,\lambda)|dt=\sqrt{\frac{2}{\pi}}\int_{-\infty}^x\frac{\lambda}{1+\lambda^2y^2}dy$$

$$\leqq\sqrt{\frac{2}{\pi}}\int_{-\infty}^\infty\frac{\lambda}{1+\lambda^2y^2}dy$$

$$=\sqrt{2\pi}.$$

Similarly $\int_0^\infty e^{-ct}|K(t,x,c,\lambda)|dx\leqq\sqrt{2\pi}$ and therefore L3 is satisfied when $a=c$. Hence the Abel kernel is an $L(a=c)$ kernel.

The Cesaro kernel of positive order

$$k(t,v,c,\lambda)=e^{(c+iv)t}\left(1-\frac{|v|}{\lambda}\right)^\alpha\chi_{(-\lambda,\lambda)}(v),\qquad\alpha>0$$

satisfies L1 and L2.

$$K(t,x,c,\lambda)=\frac{e^{ct}}{\sqrt{2\pi}}\int_{-\lambda}^\lambda\left(1-\frac{|v|}{\lambda}\right)^\alpha e^{-iv(x-t)}dv$$

$$=\sqrt{\frac{2}{\pi}}\,e^{ct}\int_0^\lambda\cos[v(x-t)]\left(1-\frac{|v|}{\lambda}\right)^\alpha dv$$

and

$$K(t,x,c,\lambda) \leqq Be^{ct}\max(\lambda,\lambda^{-\alpha}|x-t|^{-\alpha-1})$$

where B is a constant dependent on α [5, p. 30]. Therefore letting $y=x-t$,

$$\int_0^\infty e^{-at}e^{(a-c)x}|K(t,x,c,\lambda)|dt \leqq \int_{-\infty}^{x} e^{(a-c)y}B\max(\lambda,\lambda^{-\alpha}|y|^{-\alpha-1})\,dy$$

which cannot be bounded if $a<c$. We will therefore only consider the possibility $a=c$. In which case

$$\begin{aligned}\int_0^\infty e^{-ct}|K(t,x,c,\lambda)|\,dt &\leqq \int_{-\infty}^{x} B\max\left(\lambda,\lambda^{-\alpha}|y|^{-\alpha-1}\right)dy\\ &\leqq \int_{-\infty}^{\infty} B\max\left(\lambda,\lambda^{-\alpha}|y|^{-\alpha-1}\right)dy\\ &\leqq 2B\left\{\int_0^{1/\lambda} y\,dy+\int_{1/\lambda}^{\infty}\lambda^{-\alpha}y^{-\alpha-1}\,dy\right\}\\ &= 2B\left(1+\frac{1}{\alpha}\right).\end{aligned}$$

Similarly

$$\int_0^\infty e^{-ct}|K(t,x,c,\lambda)|\,dx \leqq 2B\left(1+\frac{1}{\alpha}\right)$$

and therefore L3 is satisfied when $a=c$. Hence the Cesaro kernel with $\alpha>0$ is an $L(a=c)$ kernel.

The kernel of ordinary convergence

$$k(t,v,c,\lambda)=e^{(c+iv)t}\chi_{(-\lambda,\lambda)}(v)$$

satisfies L1 and L2.

$$\begin{aligned}K(t,x,c,\lambda)&=\frac{e^{ct}}{\sqrt{2\pi}}\int_{-\lambda}^{\lambda}e^{-iv(x-t)}\,dv\\ &=\sqrt{\frac{2}{\pi}}\,e^{ct}\frac{\sin\lambda(x-t)}{x-t}.\end{aligned}$$

The kernel does not satisfy L3 even in the case $a=c$ but it does satisfy L4 for $a=c$ and $1<p\leqq 2$ since the set of transforms $\{S_\lambda\}$ forms a bounded set of transforms from $L_p(0,\infty)$ onto itself, where

$$(S_\lambda g)(t)=\frac{1}{\pi}\int_0^\infty g(x)\frac{\sin\lambda(x-t)}{x-t}\,dx,$$

for $1<p\leqq 2$ [7, p. 256]. Hence the kernel of ordinary convergence is an $L(a=c)$ kernel for $1<p\leqq 2$.

The Post–Widder kernel

$$k(t,v,c,\lambda)=\left(\frac{\lambda}{\lambda-t(c+iv)}\right)^{\lambda+1},$$

which was derived by Cooper in his work on Fourier transforms, [1, p. 292], satisfies L1 and L2.

$$K(t,x,c,\lambda)=\frac{\lambda^{\lambda+1}}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\frac{e^{-ivx}}{[\lambda-t(c+iv)]^{\lambda+1}}\,dv$$

and letting $w=\lambda-t(c+iv)$, we obtain

$$K(t,x,c,\lambda)=\frac{\lambda^{\lambda+1}}{\sqrt{2\pi}\,i|t|}e^{-\lambda x/t}e^{cx}\int_{\lambda-tc-i\infty}^{\lambda-tc+i\infty}e^{wx/t}w^{-\lambda-1}\,dw$$

$$=\frac{\sqrt{2\pi}\lambda^{\lambda+1}e^{-\lambda x/t}e^{cx}|x|^{\lambda}}{\Gamma(\lambda+1)|t|^{\lambda+1}}$$

for $x\geqq 0$ and $t\geqq 0$. L3 can only be satisfied when $a=0$, in which case

$$\int_0^{\infty}e^{-cx}|K(t,x,c,\lambda)|\,dx=\frac{\sqrt{2\pi}\lambda^{\lambda+1}}{\Gamma(\lambda+1)t^{\lambda+1}}\int_0^{\infty}e^{-\lambda x/t}x^{\lambda}\,dx$$

$$=\sqrt{2\pi}\,.$$

Similarly $\int_0^{\infty}e^{-cx}|K(t,x,c,\lambda)|\,dt=\sqrt{2\pi}$. Hence the Post–Widder kernel is an $L(0\leqq c)$ kernel.

The extended Post–Widder kernel

$$k(t,v,c,\lambda)=\left(\frac{\lambda+\theta_{\lambda}}{\lambda+\theta_{\lambda}-t(c+iv)}\right)^{\lambda+1}$$

where $\theta_{\lambda}=o(\lambda)$ as $\lambda\to\infty$, which was derived in [6, p. 82], is also an $L(0\leqq c)$ kernel.

The extended Phragmen kernel

$$k(t,v,c,\lambda)=e^{(c+iv)t}\Gamma\left(1-\frac{iv}{\lambda}\right)$$

which was derived in [6, p. 84], satisfies L1 and L2.

$$K(t,x,c,\lambda)=\frac{e^{ct}}{\sqrt{2\pi}}\int_{-\infty}^{\infty}e^{-iv(x-t)}\Gamma\left(1-\frac{iv}{\lambda}\right)dv$$

and letting $w=1-iv/\lambda$ we obtain

$$K(t,x,c,\lambda)=\frac{e^{ct}}{\sqrt{2\pi}}e^{-\lambda(x-t)}\left(\frac{\lambda}{i}\right)\int_{1-i\infty}^{1+i\infty}e^{\lambda w(x-t)}\Gamma(w)\,dw$$

$$=\sqrt{2\pi}\lambda e^{ct}e^{-\lambda(x-t)}e^{-e^{-\lambda(x-t)}}$$

by [4, p. 231].

Therefore letting $y=x-t$,

$$\int_0^{\infty}e^{-at}e^{(a-c)x}|K(t,x,c,\lambda)|\,dt=\int_{-\infty}^{x}\sqrt{2\pi}\,e^{(a-c)y}\lambda e^{-\lambda y}e^{-e^{-\lambda y}}\,dy$$

which cannot be bounded if $a<c$. We will therefore only consider the possibility $a=c$. In which case

$$\begin{aligned}\int_0^\infty e^{-ct}|K(t,x,c,\lambda)|\,dt &= \sqrt{2\pi}\int_{-\infty}^{x}\lambda e^{-\lambda y}e^{-e^{-\lambda y}}dy\\ &\leqq \sqrt{2\pi}\int_{-\infty}^{\infty}\lambda e^{-\lambda y}e^{-e^{-\lambda y}}dy\\ &= \sqrt{2\pi}\,.\end{aligned}$$

Similarly

$$\int_0^\infty e^{-ct}|K(t,x,c,\lambda)|\,dx \leqq \sqrt{2\pi}$$

and therefore L3 is satisfied when $a=c$. Hence the extended Phragmen kernel is an $L(a=c)$ kernel.

An interesting result which was noted by Cooper [2, p. 233] is that for $L(a\leqq c)$ kernels it is possible to determine the abscissa of absolute convergence of the Laplace or Laplace–Stieltjes transform, $f(w)$, by considering the behaviour of $f(w)$ on any line $w=c+iv$ in the half-plane of absolute convergence. For $L(a\leqq c)$ kernels the abscissa of absolute convergence is the infimum of the values of a for which $\|F(t,c,\lambda)e^{-at};(0,\infty)\|_1$ is bounded uniformly in λ. For $L(a=c)$ the abscissa of absolute convergence is the infimum of the values of c for which $\|F(t,c,\lambda)e^{-ct};(0,\infty)\|_1$ is bounded uniformly in λ.

REFERENCES

[1] J. L. B. Cooper, *Fourier transforms and inversion formulae for L^p functions*, Proc. London Math. Soc., 14 (1964), pp. 271–298.
[2] _____, *The representation of functions as Laplace transforms*, Math. Ann., 159 (1965), pp. 223–233.
[3] _____, *Laplace transforms of distributions*, Can. J. Math., 18 (1966), pp. 1325–1332.
[4] E. T. Copson, *Theory of Functions of a Complex Variable*, Oxford Univ. Press, Cambridge, 1935.
[5] E. C. Titchmarsh, *Fourier Integrals*, Oxford Univ. Press, Cambridge, 1962.
[6] F. J. Wilson, *Ideal inversion formulae for the Fourier transform*, this Journal, 10 (1979), pp. 80–85.
[7] A. Zygmund, *Trigonometric Series*, *Volume* II, Cambridge Univ. Press, Cambridge, 1959.

SIAM J. MATH. ANAL.
Vol. 17, No. 3, May 1986

014

ON ZEROS OF INTERPOLATING POLYNOMIALS*

ROGER W. BARNARD†, WAYNE T. FORD† AND HSING Y. WANG‡

Abstract. Polynomials to be used in interpolation of digital signals are called interpolating polynomials. They may require modification to assure convergence of their reciprocals on the unit circle.

This paper concerns discrete time windowing, which consists of scaled truncation of a series such as

$$P_N(z) \overset{\Delta}{=} 1 + \sum_{m=1}^{\infty} (z^m + z^{-m}) \operatorname{sinc} \frac{m\pi}{N}, \qquad \operatorname{sinc} x \overset{\Delta}{=} \frac{\sin x}{x},$$

where $N > 1$, to obtain an expression of the form

$$P^*_{N,L}(z) \overset{\Delta}{=} z^{L-1}\left(1 + \sum_{m=1}^{L-1} (z^m + z^{-m}) c_m \operatorname{sinc} \frac{m\pi}{N}\right).$$

We delete the asterisk to write $P_{N,L}$ when each $c_m = 1$.

The zeros of $P_{N,L}$ are shown to have unit modulus for $L \leqq N$. Examples are given to show that little can be said of the zeros of $P_{N,L}$ for $L > N$. Conditions are found to define real sequences of the form, $\{c_m : 1 \leqq m < \infty\}$, so that $P^*_{N,L}$ has no zero of unit modulus. Several standard discrete time windows are shown to define real sequences which are special cases of the conditions developed.

Introduction. Polynomials to be used in interpolation of digital signals are called interpolating polynomials. These polynomials may require modification to assure convergence of their reciprocals on the unit circle. Such modification is a principal concern of this paper.

A real function, g, defined for all values of the real independent variable time, t, is called a signal. A digital signal, γ, is a real sequence, $\{\gamma_m : -\infty < m < \infty\}$, consisting of equally spaced values or samples, $\gamma_m = g(m\Delta t)$, from the signal, g, with a time increment or sample interval, Δt. Thus, the independent variable for digital signals such as γ is sample time, $m\Delta t$, or simply sample number, m.

The signal, g, is studied in terms of its classical Fourier transform, G, as a function of real frequency, ω. The digital analog of the Fourier transform consists of the study of a sequence such as γ in terms of its Z-transform, which is defined to be the power series, Γ, having γ_m as the coefficient of z^m. Frequency's digital analog comes from evaluation of Z-transforms such as Γ on the unit circle with the negative of the θ in $z = e^{i\theta}$ referred to as frequency. If the coefficients in Γ are used without any actual evaluation of $\Gamma(z)$ or g is used without computation of G, such use is said to be in the time domain. But if $\Gamma(z)$ is used with evaluation for some z of unit modulus or G is used, such use is said to be in the frequency domain.

Signals are based on even functions in a number of applications and in this paper. This restricts digital signals to self-inversive cases meaning that $\Gamma(z) = \Gamma(z^{-1})$ for $z \neq 0$. Equivalently, γ is a symmetric sequence meaning that $\gamma_m = \gamma_{-m}$ for all m.

A second signal, f, with Fourier transform, F, poses as a filter of the signal, g, if the convolution integral, $g * f$, of g and f is considered. Of course, the Fourier transform of $g * f$ is the product of the Fourier transforms, G of g and F of f. The

*Received by the editors April 17, 1984, and in revised form October 22, 1984.
†Mathematics Department, Texas Tech University, Lubbock, Texas 79409.
‡Mathematics Department, Chinese University of Hong Kong, Hong Kong.

discrete analogy consists of the product of Z-transforms, Γ and Φ, where the latter refers to the power series with the sample, $\Phi_m = f(m\Delta t)$, taken from the filter, f, as the coefficient of z^m.

Reduction of certain frequencies is a fundamental aim in application of a filter, f, to a function, g. This can involve definition of f by the requirement that $F(\omega)$ be a constant, c, for $|\omega| < \omega_0$ but zero otherwise. If so, c can be chosen so that

$$f(t) = \operatorname{sinc} \omega_0 t, \tag{1.1}$$

where

$$\operatorname{sinc} x \overset{\Delta}{=} \frac{\sin x}{x}. \tag{1.2}$$

These equations illustrate definition of a real signal from specification of its Fourier transform. Similarly, digital signals are often defined by specification of Z-transforms.

The Fourier transform, F, of the f in (1.1) is referred to as a frequency window since it has compact support in frequency. Application of such a window to a signal, g, is known as frequency windowing. This paper concerns discrete time windowing. This consists of scaled truncation of an infinite sequence such as γ to obtain a finite sequence of the form $\{c_m \gamma_m : -L < m < L\}$ wherein the finite sequence, $\{c_m : -L < m < L\}$, is referred to as a time window.

Suppose a given digital signal, $\{b_k : -\infty < k < \infty\}$, is such that b_k is understood to correspond to the time, $kN\Delta t$, with the sample interval, $N\Delta t$, where N is a natural number such that $N > 1$. If this digital signal is to be compared with digital signals based on the smaller sample interval, Δt, the given digital signal must be interpolated to the smaller sample interval, Δt. For example, insertion of $N-1$ zeros between every b_k and b_{k+1} followed by multiplication of the Z-transform of the result by the interpolating series,

$$P_N(z) \overset{\Delta}{=} 1 + \sum_{m=1}^{\infty} (z^m + z^{-m}) \operatorname{sinc} \frac{m\pi}{N}, \tag{1.3a}$$

leads to

$$A(z) \overset{\Delta}{=} \sum_{n=-\infty}^{\infty} a_n z^n \overset{\Delta}{=} \left(\sum_{j=-\infty}^{\infty} b_j z^{jN} \right) P_N(z). \tag{1.3b}$$

Since the coefficient of z^{kN}, a_{kN}, in A comes from products of b_j and $\operatorname{sinc}(m\pi/N)$ such that $kN = jN \pm m$, it follows that $m \equiv 0 \pmod N$, $\operatorname{sinc}(m\pi/N) = 0$ for nonzero m, and $a_{kN} = b_k$. Thus, A is an interpolation of the given B.

A major purpose of this paper is to study possible alternatives to the interpolation used in (1.3a) in terms of truncation of the interpolating series in (1.3b). We consider the interpolating polynomial,

$$P_{N,L}(z) \overset{\Delta}{=} z^{L-1} \left(1 + \sum_{m=1}^{L-1} (z^m + z^{-m}) \operatorname{sinc} \frac{m\pi}{N} \right), \tag{1.4}$$

where $N > 1$.

Note that $P_{N,L}$ is a polynomial of degree $2L-2$ except that it has degree $2L-3$ and $P_{N,L}(0) = 0$ when $L \equiv 1 \pmod N$. In any case, its real coefficients imply conjugates of nonzero roots to be roots, and symmetry of coefficients implies reciprocals of

nonzero roots to be roots. Since the conjugate and the reciprocal of a root of unit modulus are equal, and the conjugate of a real root is the root itself, nonzero roots can occur in pairs. In other cases, a nonzero root, its conjugate, and their reciprocals are all different and plot as the vertices of a trapezoid in the complex plane.

We show that the zeros of $P_{N,L}$ are all of unit modulus for $L \leqq N$. Since

$$P_{N,N+1}(z) = P_{N,N}(z), \tag{1.5}$$

$P_{N,N+1}$ has a zero at the origin in addition to the zeros of unit modulus of $P_{N,N}$. We use examples to show that little can be said of the zeros of $P_{N,L}$ for $L > (N+1)$.

Conditions are then developed to define real sequences of the form, $\{c_m : 1 \leqq m < \infty\}$, so that the polynomial,

$$P^*_{N,L}(z) \overset{\Delta}{=} z^{L-1}\left(1 + \sum_{m=1}^{L-1} (z^m + z^{-m}) c_m \operatorname{sinc}\frac{m\pi}{N}\right), \tag{1.6}$$

has no zero of unit modulus. A number of standard discrete time windows are shown to define real sequences which are special cases of the conditions developed.

2. Zeros of $P_{N,L}$ for $L \leqq N$. Our study of $P_{N,L}$, for $L \leqq N$, is based on the properties of $H_{N,L}$ as defined by

$$H_{N,L}(\theta) \overset{\Delta}{=} (2\pi/N) P_{N,L}(z)/z^{L-1}\Big|_{z=e^{i\theta}}. \tag{2.1}$$

LEMMA 2.1.

$$H_{N,L}(\theta) = \int_{\theta-\pi/N}^{\theta+\pi/N} \frac{\sin[(2L-1)t/2]}{\sin(t/2)}\,dt. \tag{2.2}$$

Proof. Use the identity,

$$\begin{aligned}\int_{\theta-\pi/N}^{\theta+\pi/N} e^{imt}\,dt &= e^{im\theta}\frac{e^{im\pi/N} - e^{-im\pi/N}}{im} \\ &= \frac{\pi}{N} e^{im\theta}\frac{2\sin(m\pi/N)}{m\pi/N} = \frac{2\pi}{N} e^{im\theta}\operatorname{sinc}\frac{m\pi}{N},\end{aligned} \tag{2.3}$$

to eliminate the sinc in (1.4). Then, substitute (1.4) in (2.1) and compute

$$\begin{aligned}H_{N,L}(\theta) &= \int_{\theta-\pi/N}^{\theta+\pi/N}\left(1 + \sum_{m=1}^{L-1}(e^{imt} + e^{-imt})\right)dt \\ &= \int_{\theta-\pi/N}^{\theta+\pi/N}\left(1 + e^{it}\frac{1-e^{it(L-1)}}{1-e^{it}} + e^{-it}\frac{1-e^{-it(L-1)}}{1-e^{-it}}\right)dt \\ &= \int_{\theta-\pi/N}^{\theta+\pi/N}\left(1 + \frac{e^{it}(1-e^{it(L-1)}) - (1-e^{-it(L-1)})}{1-e^{it}}\right)dt \\ &= \int_{\theta-\pi/N}^{\theta+\pi/N}\left(1 + \frac{e^{it} - e^{itL} - 1 + e^{-it(L-1)}}{1-e^{it}}\right)dt \\ &= \int_{\theta-\pi/N}^{\theta+\pi/N}\frac{e^{i(2L-1)t/2} - e^{-i(2L-1)t/2}}{e^{it/2} - e^{-it/2}}\,dt,\end{aligned}$$

which implies (2.2). □

LEMMA 2.2. *Suppose* $0<\theta-\pi/N<\theta+\pi/N<2\pi$. *Then,*

$$\frac{dH_{N,L}(\theta)}{d\theta}=\frac{\sin[(2L-1)t/2]}{\sin(t/2)}\Bigg|_{\theta-\pi/N}^{\theta+\pi/N} \tag{2.4}$$

is zero if and only if

$$\sin(L\theta)\sin[(L-1)\pi/N]=\sin(L\pi/N)\sin[(L-1)\theta]. \tag{2.5a}$$

Proof. Differentiate (2.2) to verify (2.4). Then, set the derivative in (2.4) to zero and clear fractions to obtain

$$\begin{aligned}0&=\sin[(2L-1)(\theta+\pi/N)/2]\sin[(\theta-\pi/N)/2]\\&\quad-\sin[(2L-1)(\theta-\pi/N)/2]\sin[(\theta+\pi/N)/2]\\&\equiv\sin(L\theta)\sin[(L-1)\pi/N]\\&\quad-\sin(L\pi/N)\sin[(L-1)\theta],\end{aligned} \tag{2.5b}$$

involving a trigonometric identity which can most easily be verified by writing the sines in (2.5b) in terms of complex exponentials and combining terms on both sides to compare exponents. This completes the proof. □

LEMMA 2.3.

$$\begin{aligned}H_{N,L}(\theta)&=-(2/L)\sin(L\pi/N)\cos(L\theta)\\&\quad+2\int_0^{\pi/N}\frac{\sin(L\theta)\cos(Ls)\sin\theta-\cos(L\theta)\sin(Ls)\sin s}{\cos s-\cos\theta}\,ds\\&=2\int_0^{\pi/N}\frac{\cos(Ls)\cos[(L-1)\theta]-\cos(L\theta)\cos[(L-1)s]}{\cos s-\cos\theta}\,ds.\end{aligned} \tag{2.6}$$

Proof. Use the variable of integration, $s=t-\theta$, to write (2.2) in the form,

$$\begin{aligned}H_{N,L}(\theta)&=\int_{\theta-\pi/N}^{\theta+\pi/N}\frac{\sin(Lt)\cos(t/2)-\cos(Lt)\sin(t/2)}{\sin(t/2)}\,dt\\&=\int_{-\pi/N}^{\pi/N}\frac{\sin[L(\theta+s)]\cos[(\theta+s)/2]}{\sin[(\theta+s)/2]}-\cos[L(\theta+s)]\,ds,\end{aligned} \tag{2.7}$$

wherein the latter integrand can be written in the form,

$$\begin{aligned}\int_{-\pi/N}^{\pi/N}\cos[L(\theta+s)]\,ds&=\frac{1}{L}\sin[L(\theta+s)]\Big|_{-\pi/N}^{\pi/N}\\&=L^{-1}\{\sin[L(\theta+\pi/N)]-\sin[L(\theta-\pi/N)]\}\\&=(2/L)\sin(L\pi/N)\cos(L\theta).\end{aligned} \tag{2.8}$$

Change the integration interval from $(-\pi/N,\pi/N)$ to $(0,\pi/N)$ to write (2.7) in the form

$$
\begin{aligned}
(2.9)\quad & H_{N,L}(\theta)+(2/L)\sin(L\pi/N)\cos(L\theta) \\
&= \int_0^{\pi/N} \frac{\sin[L(\theta+s)]\cos[(\theta+s)/2]}{\sin[(\theta+s)/2]} + \frac{\sin[L(\theta-s)]\cos[(\theta-s)/2]}{\sin[(\theta-s)/2]}\,ds \\
&= \int_0^{\pi/N} \left(\frac{\cos[(\theta+s)/2]}{\sin[(\theta+s)/2]} + \frac{\cos[(\theta-s)/2]}{\sin[(\theta-s)/2]} \right) \sin(L\theta)\cos(Ls) \\
&\quad + \left(\frac{\cos[(\theta+s)/2]}{\sin[(\theta+s)/2]} - \frac{\cos[(\theta-s)/2]}{\sin[(\theta-s)/2]} \right) \cos(L\theta)\sin(Ls)\,ds \\
&= \int_0^{\pi/N} \frac{\sin(L\theta)\cos(Ls)\sin\theta - \cos(L\theta)\sin(Ls)\sin s}{\sin[(\theta+s)/2]\sin[(\theta-s)/2]}\,ds,
\end{aligned}
$$

which implies the first equality in (2.6). Observe that

$$
\sin(L\pi/N)\cos(L\theta) = L\int_0^{\pi/N} \frac{\cos(Ls)\cos(L\theta)[\cos s-\cos\theta]}{\cos s-\cos\theta}\,ds,
$$

shows that $H_{N,L}(\theta)$ is given by an integral in which the numerator of the integrand has the form,

$$
\begin{aligned}
&-\cos(Ls)\cos(L\theta)[\cos s-\cos\theta]+\sin(L\theta)\cos(Ls)\sin\theta-\cos(L\theta)\sin(Ls)\sin s \\
&\quad=\cos(Ls)[\cos(L\theta)\cos\theta+\sin(L\theta)\sin\theta]-\cos(L\theta)[\cos(Ls)\cos s+\sin(Ls)\sin s],
\end{aligned}
$$

which implies the remaining equality in (2.6). □

THEOREM 2.1. *The zeros of $P_{N,L}$ have unit modulus for $L \leqq N$.*

Proof. Set L to N in (2.5) to observe that

$$
(2.10)\qquad \frac{dH_{N,N}(\theta)}{d\theta}=0 \quad \text{iff} \quad \sin(N\theta)=0 \quad \text{iff} \quad \theta=k\pi/N.
$$

Although $\theta=k\pi/L$ is not a zero of the derivative of $H_{N,L}$, we use it to write the first equality in (2.6) in the form,

$$
\begin{aligned}
(2.11)\qquad H_{N,L}(k\pi/L) &= (-1)^{k+1}(2/L)\sin(L\pi/N) \\
&\quad +2(-1)^{k+1}\int_0^{\pi/N} \frac{\sin(Ls)\sin s}{\cos s-\cos(k\pi/L)}\,ds.
\end{aligned}
$$

Since the above integrand is positive on $(0,\pi/N)$ if $L \leqq Nk$, the integral in (2.11) is positive. Thus, $H_{N,L}(k\pi/L)$ has $(L-1)$ changes of sign as k counts from 1 to L, $P_{N,L}(z)$ has $(L-1)$ zeros in the upper half of the unit circle, $(L-1)$ conjugate zeros in the lower half of the unit circle, and the proof is complete. □

3. Zeros of $P_{N,L}$ for $L>N$. Several examples are given to show that Theorem 2.1 cannot be extended to cover $L>N$. The most trivial example,

$$
(3.1)\qquad P_{N,N+1}=zP_{N,N},
$$

has a zero at the origin in addition to the zeros of unit modulus of $P_{N,N}$.

We discuss $P_{N,L}$, for $L>N+1$, in terms of $H_{N,L}$, given in (2.1), and $F_{N,L}$, defined by

$$F_{N,L}(x) \overset{\Delta}{=} H_{N,L}(\theta)\Big|_{\cos\theta = x}. \tag{3.2}$$

Observe that each zero of $F_{N,L}$ in the interval, $(-1,1)$, implies two zeros of $H_{N,L}$, which implies two zeros of unit modulus of $P_{N,L}$.

The example,

$$\begin{aligned} H_{2,4}(\theta) &= \pi + 4\sum_{m=1}^{3} \frac{1}{m}\sin\frac{m\pi}{2}\cos m\theta \\ &= \pi + 4\cos\theta - \frac{4}{3}\cos 3\theta \\ &= \pi + 8\cos\theta - \frac{16}{3}\cos^3\theta, \end{aligned} \tag{3.3a}$$

defines

$$F_{2,4}(x) = \pi + 8x - \frac{16}{3}x^3, \qquad x = \cos\theta, \tag{3.3b}$$

which has three real roots, two in $(-1,0)$ and one in $(1,\infty)$, since it is positive at $x=-1$, $x=0$, and $x=1$, but it is negative for $x=-\frac{1}{2}$ and for large positive x. Its two roots in $(-1,0)$ force four roots of unit modulus on the polynomial,

$$\begin{aligned} P_{2,4}(z) &= \sum_{m=0}^{6} z^m \operatorname{sinc}\frac{(m-3)\pi}{2} \\ &= \frac{1}{3\pi}(-2+6z^2+3\pi z^3+6z^4-2z^6). \end{aligned} \tag{3.3c}$$

Since $P_{2,4}(0)<0$ and $P_{2,4}(1)>0$, the two remaining roots consist of one root in $(0,1)$ and its reciprocal in $(1,\infty)$.

The second example,

$$\begin{aligned} H_{2,6}(\theta) &= \pi + 4\sum_{m=1}^{5} \frac{1}{m}\sin\frac{m\pi}{2}\cos m\theta \\ &= \pi + 4\cos\theta - \frac{4}{3}\cos 3\theta + \frac{4}{5}\cos 5\theta \\ &= \pi + 4\left(3\cos\theta - \frac{16}{3}\cos^3\theta + \frac{16}{5}\cos^5\theta\right) \end{aligned} \tag{3.4a}$$

defines

$$F_{2,6}(x) = \pi + 4\left(3x - \frac{16}{3}x^3 + \frac{16}{5}x^5\right), \qquad x = \cos\theta, \tag{3.4b}$$

which has three real roots in $(-1,0)$ and two complex roots, since it has the same sign as x as $|x|\to\infty$, a positive maximum at $-\sqrt{3}/2$, a negative minimum at $-1/2$, a positive maximum at $1/2$, and a positive minimum at $\sqrt{3}/2$. The three roots of $F_{2,6}$ in

$(-1,0)$ force six roots of unit modulus on the polynomial,

$$P_{2,6}(z)=\sum_{m=0}^{10} z^m \operatorname{sinc}\frac{(m-5)\pi}{2} \tag{3.4c}$$

$$=\frac{2}{15\pi}\left(3-5z^2+15z^4+\frac{15\pi}{2}z^5+15z^6-5z^8+3z^{10}\right)$$

$$=\left(1+\frac{2}{15\pi}B(z)\right)z^5, \quad \text{where } B(z)=Q(z^{-1})+Q(z)$$

with

$$Q(z)=3z^5-5z^3+15z. \tag{3.5a}$$

Since

$$\frac{dQ}{dz}=\frac{15(z^6+1)}{z^2+1} \quad \text{and} \quad \frac{dB}{dz}=\frac{15(z^{12}-1)}{z^6(z^2+1)}, \tag{3.5b}$$

it follows that $xB(x)>0$ for nonzero real x,

$$\inf\{|B(x)|: x \text{ real}\}=|B(-1)|=26,$$

$$\sup\{1+2B(x)/(15\pi): x<0\}=1-52/(15\pi)<0,$$

and $P_{2,6}$ has no real roots. Thus, its remaining four roots must form the vertices of a trapezoid in the complex plane.

4. Zeros of $P^*_{N,L}$. We seek real sequences of the form, $\{c_m: 1\leqq m<\infty\}$, such that polynomials defined by (1.6) have no roots in $\{z:|z|=1\}$. The search will be based on using the same real sequences in defining the polynomials,

$$Q_{N,L}(z)\overset{\Delta}{=}1+2\sum_{m=1}^{L-1} c_m z^m \operatorname{sinc}\frac{m\pi}{N}, \tag{4.1}$$

which will then be such that

$$\operatorname{Re}\left[Q_{N,L}(e^{i\theta})\right]\cos(L-1)\theta=\operatorname{Re}\left[P^*_{N,L}(e^{i\theta})\right]. \tag{4.2}$$

DEFINITION 4.1. R denotes the class of functions which are analytic and of positive real part on $\{z:|z|<1\}$.

First, take L to be infinite in (4.1), set $c_m=1$ for all m, and denote the result by

$$w(\zeta)\overset{\Delta}{=}1+2\sum_{m=1}^{\infty}\zeta^m \operatorname{sinc}\frac{m\pi}{N}, \tag{4.3}$$

which will be shown below to lie in R. A classical result will then be used to develop conditions on $\{c_m: 1\leqq m<\infty\}$ to imply $Q_{N,L}\in R$. Another classical result will then be used to show that the same conditions imply that $P^*_{N,L}$, as given in (1.6), has no zeros of unit modulus. Alternative criteria for determining whether a given $\{c_m: 1\leqq m<\infty\}$ has the desired properties will then be shown.

LEMMA 4.1. *The function, w, maps $\{\zeta:|\zeta|<1\}$ onto the vertical strip bounded by $\{w:\operatorname{Re}w=0\}\cup\{w:\operatorname{Re}w=N\}$. Thus, $w\in R$.*

Proof. Compute

$$w(\zeta)=1+\frac{N}{i\pi}\sum_{m=1}^{\infty}\frac{\zeta^m}{m}(e^{im\pi/N}-e^{-im\pi/N}) \tag{4.4a}$$

$$=1+\frac{N}{i\pi}\left(\sum_{m=1}^{\infty}\frac{\zeta^m e^{im\pi/N}}{m}-\sum_{m=1}^{\infty}\frac{\zeta^m e^{-im\pi/N}}{m}\right)$$

$$=1+\frac{N}{i\pi}\left(-\ln(1-\zeta e^{i\pi/N})+\ln(1-\zeta e^{-i\pi/N})\right)$$

$$=1+\frac{N}{i\pi}\ln\frac{1-\zeta e^{-i\pi/N}}{1-\zeta e^{i\pi/N}}.$$

Since w is analytic for $|\zeta|<1$ and $w(0)=1$, it suffices to consider $w(\zeta)$ for $|z|=1$. Compute

$$\operatorname{Re} w(e^{i\Phi})=1+\operatorname{Re}\left(\frac{N}{i\pi}\ln\frac{1-e^{i\Phi-i\pi/N}}{1-e^{i\Phi+i\pi/N}}\right) \tag{4.4b}$$

$$=1+\frac{N}{\pi}\operatorname{Im}\ln\frac{e^{i(\Phi-\pi/N)/2}\left(e^{-i(\Phi-\pi/N)/2}-e^{i(\Phi-\pi/N)/2}\right)}{e^{i(\Phi+\pi/N)/2}\left(e^{-i(\Phi+\pi/N)/2}-e^{i(\Phi+\pi/N)/2}\right)}$$

$$=1+\frac{N}{\pi}\operatorname{Im}\ln\left(e^{-i\pi/N}\frac{\sin[(\Phi-\pi/N)/2]}{\sin[(\Phi+\pi/N)/2]}\right)$$

$$=\begin{cases}1+\frac{N}{\pi}\left(-\frac{\pi}{N}+\pi\right)\\ 1+\frac{N}{\pi}\left(-\frac{\pi}{N}\right)\end{cases}=\begin{cases}N, & \Phi\in(\pi/N,\ 2\pi-\pi/N),\\ 0, & \Phi\notin(\pi/N,\ 2\pi-\pi/N)\end{cases}. \qquad \square$$

LEMMA 4.2. *Let L be an integer exceeding unity. Let r_L denote the unique positive root of*

$$2r^L+r-1=0. \tag{4.5}$$

Then, $0<r_L<1$ and $r_L<r_{L+1}$. Also, $r_L>1-(2/L)\ln L$, and $r_L\to 1$ as $L\to\infty$. Adopt the definitions,

$$f(\zeta)\overset{\Delta}{=}\sum_{n=0}^{\infty}a_n\zeta^n \quad \text{and} \quad s_L(\zeta)\overset{\Delta}{=}\sum_{n=0}^{L-1}a_n\zeta^n. \tag{4.6}$$

If $f\in R$, then $\operatorname{Re}[s_L(\zeta)]>0$ on $\{\zeta:|\zeta|<r_L\}$. Moreover, the example using $a_0=1$ and $a_n=2$ for $n>0$ shows that r_L cannot be increased in the conclusion.

Proof. [2, p. 523]. $\square$

LEMMA 4.3. *Let $L>1$, $r\in(0,r_L)$ with r_L defined by* (4.5), *and f as denoted in* (4.6). *If $f\in R$, then*

$$\operatorname{Re}\left(\sum_{m=0}^{L-1}a_m r^m z^m\right)>0 \quad \text{on } \{z:|z|\leqq 1\}. \tag{4.7}$$

Proof. Set $\zeta=rz$ in partial sums of f as denoted in (4.6), and note that $|z|\leqq 1$ if and only if $|\zeta|\leqq r<r_L$. Then, Lemma 4.1 implies (4.7). $\square$

LEMMA 4.4. *Let* $\{b_n : 0 \leqq n < \infty\} \in l_1$ *such that* $b_0 \neq 0$, *and let* f *be denoted as in* (4.6). *Then*,

$$\operatorname{Re}\left(\frac{1}{b_0}\sum_{n=0}^{\infty} b_n a_n\right) \geqq 0 \tag{4.8}$$

for all $f \in R$ *if and only if*

$$\operatorname{Re}\left(\frac{1}{b_0}\sum_{n=0}^{\infty} b_n z^n\right) \geqq \frac{1}{2} \quad \text{on } \{z : |z| = 1\}. \tag{4.9}$$

Proof. [2, pp. 517–518]. □

LEMMA 4.5. *Lemma* 4.4 *remains valid with* (4.8) *replaced by*

$$\operatorname{Re}\left(\frac{1}{b_0}\sum_{n=0}^{\infty} b_n a_n z^n\right) \geqq 0 \quad \text{for } |z| < 1. \tag{4.10}$$

Proof. Clearly, validity of (4.10) for all $f \in R$ implies validity of (4.8) for all $f \in R$. It remains to show that Lemma 4.4 implies validity of (4.10) for all $f \in R$. Fix ζ with $0 < |\zeta| < 1$ and let $f \in R$. Then,

$$g_\zeta(z) \triangleq \sum_{n=0}^{\infty}\left(\frac{a_n \zeta^n}{|\zeta|^n}\right) z^n \tag{4.11}$$

satisfies $g_\zeta \in R$. Thus, (4.8) can be written in the form,

$$\operatorname{Re}\left(\frac{1}{b_0}\sum_{n=0}^{\infty}\frac{b_n a_n \zeta^n}{|\zeta|^n}\right) \geqq 0 \quad \text{for } |\zeta| \leqq 1, \tag{4.12}$$

wherein setting $|\zeta| = 1$ merely reduces (4.12) to (4.8). Let $z = \zeta/|\zeta|$ to write (4.12) in the form,

$$\operatorname{Re}\left(\frac{1}{b_0}\sum_{n=0}^{\infty} b_n a_n z^n\right) \geqq 0 \quad \text{for } |z| = 1. \tag{4.13}$$

Since $f \in R$ implies $\operatorname{Re}(a_0) \geqq 0$, (4.10) applies for $z = 0$, which combines with (4.13) to imply (4.10). □

THEOREM 4.1. *Suppose* $\{b_n : 0 \leqq n < \infty,\ b_0 = 1\}$ *is a real sequence satisfying* (4.9). *Let*

$$c_m \equiv r^m b_m, \quad \text{where } 0 \leqq r < r_L \tag{4.14}$$

with r_L *being the positive root of* (4.5). *Then*, $P^*_{N,L}$, *defined by* (1.6), *has no zeros in* $\{z : |z| = 1\}$.

Proof. Lemma 4.1 shows $w \in R$. Apply Lemma 4.5 to w to show that $f \in R$, where

$$f(\zeta) \triangleq 1 + 2\sum_{m=1}^{\infty} b_m \zeta^m \operatorname{sinc}\frac{m\pi}{N}. \tag{4.15}$$

Apply Lemma 4.3 to show that

$$Q_{N,L}(z) \triangleq 1 + 2\sum_{m=1}^{L-1} b_m r^m z^m \operatorname{sinc}\frac{m\pi}{N} \tag{4.16}$$

has positive real part on $\{z:|z|\leqq 1\}$. Since (4.14) shows (4.16) to be the same as (4.1), (4.2) implies the desired result. □

Classical tests to determine whether a given finite sequence, $\{c_m: 0<m<L\}$, can be used to define a window implying the results in Theorem 4.1 are given below.

THEOREM 4.2. *A sequence,* $\{b_m: 0\leqq m<L,\ b_0=1\}$, *initiates some infinite sequence,* $\{b_m: 0\leqq m<\infty,\ b_0=1\}$, *such that*

$$\operatorname{Re}\left(1+\sum_{m=1}^{\infty} b_m z^m\right)>\frac{1}{2} \quad \text{for } |z|<1 \tag{4.17}$$

if and only if

$$\begin{vmatrix} 1 & b_1 & \cdot & & & b_{k-1} & b_k \\ \bar{b}_1 & 1 & b_1 & \cdot & & & b_{k-1} \\ \cdot & \cdot & \cdot & \cdot & \cdot & & \\ & \cdot & \cdot & \cdot & \cdot & \cdot & \\ & & \cdot & \cdot & \cdot & \cdot & \cdot \\ \bar{b}_{k-1} & & & \cdot & \bar{b}_1 & 1 & b_1 \\ \bar{b}_k & \bar{b}_{k-1} & & & \cdot & \bar{b}_1 & 1 \end{vmatrix} \geqq 0 \tag{4.18}$$

for $0<k<L$ [6]. *Moreover,* (4.17) *is equivalent to the existence of a probability measure,* Ψ, *on* $[0,2\pi]$ *such that*

$$b_m=\frac{1}{2\pi}\int_0^{2\pi} e^{im\theta}\,d\Psi(\theta), \qquad 0\leqq m<\infty. \tag{4.19}$$

Proof. [5] for (4.18) and [7] for (4.19). □

THEOREM 4.3. *If the real sequence,* $\{b_m: 0\leqq m<L,\ b_0=1\}$, *is such that* (4.18) *is satisfied for* $0<k<L$, *let*

$$c_m=b_m\left(1-\frac{2\log L}{L}\right)^m \tag{4.20}$$

define the coefficients in (1.6). *Then* $P^*_{N,L}$ *has no zero of unit modulus.*

Proof. The inequality above (4.6) shows that (4.20) defines c_m satisfying the hypotheses of Theorem 4.1. □

5. Windows. The generalized Hamming window [4] is a standard parameterized time window which defines real sequences satisfying the conditions in Theorem 4.2. This window is defined by

$$\begin{aligned} &b_m=\alpha+(1-\alpha)\cos\frac{2\pi m}{K-1}=\alpha+(1-\alpha)\cos\frac{\pi m}{J} \\ &\text{for } K=2J+1 \quad \text{and} \quad -J=-\frac{K-1}{2}\leqq m\leqq\frac{K-1}{2}=J, \end{aligned} \tag{5.1a}$$

or

$$\begin{aligned} &b_m=\alpha+(1-\alpha)\cos\frac{2\pi(2m+1)}{2(K-1)}=\alpha+(1-\alpha)\cos\frac{\pi(2m+1)}{2J-1} \\ &\text{for } K=2J \quad \text{and} \quad -J=-\frac{K}{2}\leqq m\leqq\frac{K}{2}-1=J-1, \end{aligned} \tag{5.1b}$$

wherein α ordinarily lies in $[\frac{1}{2},1)$. The generalized Hamming window is known as the Hamming window if $\alpha=0.54$ and as the Hanning window if α is one-half.

The even case, with $K=2J$, is discarded here for lack of symmetry. Then, the generalized Hamming window becomes

$$b_m=\alpha+\frac{1-\alpha}{2}\left(e^{i\pi m/J}+e^{-i\pi m/J}\right) \tag{5.2}$$

$$=\alpha+(1-\alpha)\cos\frac{\pi m}{J} \quad \text{for } -J\leqq m\leqq J.$$

THEOREM 5.1. *Use b_m given by* (5.2) *with $J=N-1>0$ and $0<\alpha<1$. Then, $\{b_m: 0\leqq m<N-1,\ b_0=1\}$ initiates the infinite sequence, $\{b_m: 0\leqq m<\infty,\ b_0=1\}$, satisfying* (4.17) *if b_m is defined by* (5.2) *for all m.*

Proof. Compute

$$\sum_{m=0}^{\infty} b_m z^m=\frac{\alpha}{1-z}+\frac{1-\alpha}{2}\left(\frac{1}{1-ze^{i\pi/N}}+\frac{1}{1-ze^{-i\pi/N}}\right). \tag{5.3}$$

Since $(1-z)^{-1}$ maps the unit disc onto $\{w:\mathrm{Re}(w)>1/2\}$, and the bracket in (5.3) is a sum of compositions of $(1-z)^{-1}$ and rotations of the unit disc,

$$\mathrm{Re}\left(\sum_{m=0}^{\infty} b_m z^m\right)>\frac{\alpha}{2}+\frac{1-\alpha}{2}=\frac{1}{2}, \qquad |z|<1. \tag{5.4}$$

□

REFERENCES

[1] W. T. FORD, *Optimum mixed delay spiking filters*, Geophysics, 43 (1978), pp. 125–132.

[2] G. M. GOLUSIN, *Geometric Theory of Functions of a Complex Variable*, AMS Transl. Math. Mono., vol. 26, American Mathematical Society, Providence, RI, 1969.

[3] C. POMMERENKE, *Univalent Functions*, Vanderhoeck and Ruprecht, Gottingen, 1975.

[4] L. R. RABINER AND B. GOLD, *Theory and Application of Digital Signal Processing*, Prentice-Hall, Englewood Cliffs, NJ, 1975.

[5] W. ROGOSINSKI, *Uber positive harmonische Entwicklungen und typisch-reelle Potenzreihen*, Math. Z., 35 (1932), pp. 93–121.

[6] I. SCHUR, *Uber Potenzreihen, die im Innern des Einheitskreises beschrankt sind*, J. Reine Angew. Math., 147 (1917), pp. 205–232.

[7] H. S. WILF, *Subordinating factor sequences for convex maps of the unit circle*, Proc. AMS, 12 (1961), pp. 689–693.

SIAM J. MATH. ANAL.
Vol. 17, No. 3, May 1986

015

ASYMPTOTICS FOR THE GREATEST ZEROS OF ORTHOGONAL POLYNOMIALS*

ATTILA MÁTÉ†, PAUL NEVAI‡ AND VILMOS TOTIK§

Abstract. Asymptotics for the greatest zeros of symmetric orthogonal polynomials is investigated in terms of the asymptotic behavior of the recursion coefficients.

AMS(MOS) subject classification. Primary 42C05

Key words. orthogonal polynomials, zeros of polynomials

Let $d\alpha$ be a positive measure on the real line such that $\operatorname{supp}(d\alpha)$ is an infinite set and all the moments of $d\alpha$ are finite. In addition, we will also assume that all the odd moments of $d\alpha$ vanish. The corresponding orthogonal polynomials are denoted by $p_n(d\alpha)$, $n=0,1,\cdots$, where $p_n(d\alpha,x)=\gamma_n(d\alpha)x^n+\cdots$. The object of this paper is to investigate the asymptotic behavior of the greatest zero $X_n(d\alpha)$ of $p_n(d\alpha)$ in terms of the recursion coefficients $a_n(d\alpha)$ in the three-term recurrence formula

$$xp_n=a_{n+1}p_{n+1}+a_np_{n-1}$$

where $a_0(d\alpha)=0$ and

$$a_n(d\alpha)=\gamma_{n-1}(d\alpha)/\gamma_n(d\alpha), \qquad n=1,2,\cdots.$$

We will be concerned with the case when $\operatorname{supp}(d\alpha)$ is unbounded, and it is well known that this is equivalent to the unboundedness of the sequence $\{X_n(d\alpha)\}$, and the latter is equivalent to the unboundedness of $\{a_n(d\alpha)\}$. An example of particular significance is the case of the Hermite polynomials, which are orthonormal with respect to

$$dH(x)=\exp(-x^2)\,dx.$$

In this case

$$a_n(dH)=\sqrt{n/2} \tag{1}$$

and

$$\lim_{n\to\infty} X_n(dH)n^{-1/2}=\sqrt{2}$$

(see e.g. Szegö [20, p. 106 and p. 132]). G. Freud [5] considered

$$d\alpha_m(x)=\exp(-|x|^m)\,dx, \qquad m>0 \tag{2}$$

and proved

$$\lim_{n\to\infty} X_n(d\alpha_m)n^{-1/m}=\left[\sqrt{\pi}\,\frac{\Gamma(m/2)}{\Gamma((m+1)/2)}\right]^{1/m} \tag{3}$$

* Received by the editors August 9, 1984, and in final revised form November 8, 1984.

† Department of Mathematics, Brooklyn College of the City University of New York, Brooklyn, New York 11210. The work of this author was supported in part by the National Science Foundation under grant MCS 8100673, and by the PCS-CUNY Research Award Program of the City University of New York under grant 662043.

‡ Department of Mathematics, Ohio State University, Columbus, Ohio 43210. The work of this author was supported by the National Science Foundation under grant MCS-83-00882.

§ Bolyai Institute, University of Szeged, 6720 Szeged, Hungary. The work of this author was performed while he was visiting Ohio State University.

for $m=4$ and $m=6$. In this same paper Freud conjectured that (3) holds for every $m>0$. Freud's conjecture was proved by Rahmanov [19] for $m>1$. Another conjecture by Freud [4] states that if $d\alpha_m$ is given by (2) then

$$\lim_{n\to\infty} a_n(d\alpha_m)n^{-1/m}=\frac{1}{2}\left[\sqrt{\pi}\,\frac{\Gamma(m/2)}{\Gamma((m+1)/2)}\right]^{1/m}. \tag{4}$$

Freud [4] proved (4) for $m=4$ and $m=6$, whereas for all even integral values of m (4) was recently proved by Magnus [8]. There is an intimate connection between $X_n(d\alpha)$ and $a_n(d\alpha)$, namely

$$X_n(d\alpha)=2\max_{j_k\geq 0}\frac{\sum_{k=1}^{n}a_{n-k}(d\alpha)\,j_k j_{k+1}}{\sum_{k=1}^{n}j_k^2}, \qquad j_k \text{ real}. \tag{5}$$

This relationship was proved by Freud [5] and it is a slight improvement of a result by Chebyshev [19, p. 188]. Formula (5) arises from the Rayleigh quotient of the relevant Jacobi matrix. On the basis of (5) it is an easy exercise [5] to show that (4) implies (3). It was shown by Lew–Quarles [6] ($m=4$), Máté–Nevai [9] ($m=6$) and Máté–Nevai–Zaslavsky [10] (m even) that

$$a_n(d\alpha_m)n^{-1/m}=\frac{1}{2}\left[\sqrt{\pi}\,\frac{\Gamma(m/2)}{\Gamma((m+1)/2)}\right]^{1/m}+O(n^{-2})$$

and thus one may expect to obtain asymptotics better than (3). In fact, for the Hermite case it is known that

$$X_n(dH)n^{-1/2}=\sqrt{2}-2^{-1/2}3^{-1/3}i_1n^{-2/3}+o(n^{-2/3}), \tag{6}$$

where i_1 is the least zero of Airy's function defined as the unique solution of

$$z''+xz/3=0$$

which remains bounded as $x\to-\infty$ (see e.g. [17, p. 408] and [20, p. 132]), and this may be used to prove the following result.

THEOREM. *Suppose that all the odd moments of $d\alpha$ vanish and the recursion coefficients $a_n(d\alpha)$ satisfy*

$$a_n(d\alpha)=cn^{\delta}\left[1+o(n^{-2/3})\right], \tag{7}$$

where $c>0$ and $\delta>0$ are independent of n. Then

$$X_n(d\alpha)n^{-\delta}=2c-c3^{-1/3}(2\delta)^{2/3}i_1n^{-2/3}+o(n^{-2/3}) \tag{8}$$

where i_1 is the smallest zero of Airy's function ($i_1=3.3721\cdots$).

Remark. The underlying idea in the proof of (8) is that the problem of estimating the largest zero $X_n(d\alpha)$ of the orthogonal polynomials associated with $d\alpha$ is transformed into another problem which is essentially equivalent to estimating $X_n(H)$ and then (6) is applied. In view of the importance of the asymptotic formula (6) we recall here the outline of its proof whereas for details we refer the reader to [17] and [20]. The Hermite polynomials h_n satisfy the differential equation

$$z''+x(2\sqrt{2n+1}-x)z=0,$$

where

$$z(x)=h_n(\sqrt{2n+1}-x)\exp\bigl(-(\sqrt{2n+1}-x)^2/2\bigr).$$

This equation is almost a linearly transformed form of Airy's equation

$$z''+xz/3=0.$$

Based on this observation, Sturm's comparison theorem applied to these two differential equations yields (6).

Proof of the theorem. In order to simplify notation, we will write X_n and a_n instead of $X_n(d\alpha)$ and $a_n(d\alpha)$ respectively. First we will estimate X_n from above. Let us fix Δ such that $0<\Delta<1$ and choose $\varepsilon=\varepsilon(\Delta,\delta)$ so that $0<\varepsilon<1/2$, $\varepsilon<1/(4\delta\Delta)$ and

$$(1-x)^{\delta}\le(1-2\delta\Delta x)^{1/2},\qquad 0\le x\le 2\varepsilon, \tag{9}$$

holds. For given n we can apply (5) to find $j_k^*=j_k^*(n)$, $k=1,2,\cdots,n$, such that $j_k^*\ge 0$ and

$$\frac{1}{2}X_n=\frac{\sum_{k=1}^{n}a_{n-k}j_k^*j_{k+1}^*}{\sum_{k=1}^{n}(j_k^*)^2}.$$

Since[1]

$$\sum_{k=[\varepsilon n]}^{[2\varepsilon n]}j_k^*j_{k+1}^*\le\sum_{k=1}^{n}(j_k^*)^2,$$

there exists ν such that $[\varepsilon n]\le\nu\le[2\varepsilon n]$ and

$$\frac{j_\nu^*j_{\nu+1}^*}{\sum_{k=1}^{n}(j_k^*)^2}\le\frac{1}{\varepsilon n-1}.$$

Hence

$$\frac{1}{2}X_n\le\frac{a_{n-\nu}}{\varepsilon n-1}+\max\left\{\frac{\sum_{k=1}^{\nu-1}a_{n-k}j_k^*j_{k+1}^*}{\sum_{k=1}^{\nu}(j_k^*)^2},\frac{\sum_{k=\nu+1}^{n}a_{n-k}j_k^*j_{k+1}^*}{\sum_{k=\nu+1}^{n}(j_k^*)^2}\right\},$$

and consequently

$$\frac{1}{2}X_n\le\frac{\max_{1\le k\le n}a_{n-k}}{\varepsilon n-1}+\max\left\{\max_{j_k\ge 0}\frac{\sum_{k=1}^{[2\varepsilon n]-1}a_{n-k}j_kj_{k+1}}{\sum_{k=1}^{[2\varepsilon n]}j_k^2},\max_{[\varepsilon n]\le k\le n}a_{n-k}\right\}. \tag{10}$$

Now we will show that for sufficiently large values of n the inequality

$$\max_{[\varepsilon n]\le k\le n}a_{n-k}\le\max_{j_k\ge 0}\frac{\sum_{k=1}^{[2\varepsilon n]-1}a_{n-k}j_kj_{k+1}}{\sum_{k=1}^{[2\varepsilon n]}j_k^2} \tag{11}$$

holds. If we assume that there exists a sequence $n_1<n_2<\cdots$ such that (11) is not true for $n=n_l$ then by (10)

$$\frac{1}{2}X_n\le\frac{\max_{1\le k\le n}a_{n-k}}{\varepsilon n-1}+\max_{[\varepsilon n]\le k\le n}a_{n-k},\qquad n=n_l,$$

[1] $[z]$ denotes the integer part of z.

and thus by (7)

$$\limsup_{l\to\infty} X_{n_l}(n_l)^{-\delta} \le 2c(1-\varepsilon)^{\delta} < 2c,$$

and this contradicts

$$\lim_{n\to\infty} X_n n^{-\delta} = 2c,$$

which was proved by Freud [5, Thm. 7] whenever (7) holds. Having proved (11), we obtain from (7) and (10) the inequality

$$\frac{1}{2} X_n n^{-\delta} \le c \max_{j_k \ge 0} \frac{\sum_{k=1}^{[2\varepsilon n]-1} (1-k/n)^{\delta} j_k j_{k+1}}{\sum_{k=1}^{[2\varepsilon n]} j_k^2} + o(n^{-2/3}), \tag{12}$$

so that by (9)

$$\frac{1}{2} X_n n^{-\delta} \le c \max_{j_k \ge 0} \frac{\sum_{k=1}^{[2\varepsilon n]-1} (1-2\delta\Delta k/n)^{1/2} j_k j_{k+1}}{\sum_{k=1}^{[2\varepsilon n]} j_k^2} + o(n^{-2/3}).$$

Introducing the notation $N = [n/(2\delta\Delta)]$, we can rewrite the previous inequality as

$$\frac{1}{2} X_n n^{-\delta} \le c\sqrt{\frac{2}{N}} \max_{j_k \ge 0} \frac{\sum_{k=1}^{[2\varepsilon n]-1} \sqrt{(N-k)/2}\, j_k j_{k+1}}{\sum_{k=1}^{[2\varepsilon n]} j_k^2} + o(n^{-2/3}).$$

Taking (1) and (5) into consideration, we obtain

$$X_n n^{-\delta} \le c\sqrt{\frac{2}{N}}\, X_N(dH) + o(n^{-2/3}),$$

and thus (6) yields

$$X_n n^{-\delta} \le 2c - c3^{-1/3} i_1 N^{-2/3} + o(n^{-2/3}).$$

Hence

$$X_n n^{-\delta} \le 2c - c3^{-1/3}(2\delta)^{2/3} i_1 n^{-2/3}\Delta^{2/3} + o(n^{-2/3}),$$

and since $0 < \Delta < 1$ is arbitrary we can let Δ tend to 1 to obtain the upper estimate

$$X_n n^{-\delta} \le 2c - c3^{-1/3}(2\delta)^{2/3} i_1 n^{-2/3} + o(n^{-2/3}). \tag{13}$$

Estimating X_n from below can be achieved along lines similar to the previous arguments. We pick $\Delta > 1$ and choose $\varepsilon = \varepsilon(\Delta, \delta)$ such that $0 < \varepsilon < 1/2$, $\varepsilon < \delta\Delta$, $\varepsilon < 1/(4\delta\Delta)$ and

$$(1-2\delta\Delta x)^{1/2} \le (1-x)^{\delta}, \qquad 0 \le x \le 2\varepsilon/(\delta\Delta). \tag{14}$$

Let $N = [n/(2\delta\Delta)]$. Then we can apply (12) with $d\alpha = dH$ and $n = N$ to obtain

$$\frac{1}{2} X_N(dH) N^{-1/2} \le \frac{1}{\sqrt{2}} \max_{j_k \ge 0} \frac{\sum_{k=1}^{[2\varepsilon N]-1} (1-k/N)^{1/2} j_k j_{k+1}}{\sum_{k=1}^{[2\varepsilon N]} j_k^2} + o(N^{-2/3}).$$

Since $N = [n/(2\delta\Delta)]$ and (14) holds, we get

$$\frac{1}{2} X_N(dH) N^{-1/2} \le \frac{n^{-\delta}}{\sqrt{2}} \max_{j_k \ge 0} \frac{\sum_{k=1}^{[n\varepsilon/(\delta\Delta)]-1} (n-k)^{\delta} j_k j_{k+1}}{\sum_{k=1}^{[n\varepsilon/(\delta\Delta)]} j_k^2} + o(N^{-2/3}),$$

and therefore by (5) and (7)

$$X_N(dH)N^{-1/2} \leq \frac{n^{-\delta}}{c\sqrt{2}} X_n + o(n^{-2/3}).$$

Now the inequality

$$X_n n^{-\delta} \geq 2c - c3^{-1/3}(2\delta)^{2/3} i_1 n^{-2/3} + o(n^{-2/3}) \tag{15}$$

follows immediately from (6) by letting $\Delta \to 1$. In view of (13) and (15) the Theorem has completely been proved.

There exist several orthogonal polynomial systems whose recursion coefficients satisfy (7) and thus (8) can be applied to find asymptotics for $X_n(d\alpha)$.

The associated Pollaczek polynomials are orthogonal with respect to

$$d\alpha(x) = |\Gamma(\lambda+\gamma+ix)|^2 |{}_2F_1(1-\lambda+ix, \gamma; \gamma+\lambda+ix; -1)|^{-2} dx, \qquad -\infty < x < \infty,$$

where either $2\lambda+\gamma > 0$, $\gamma \geq 0$ or $2\lambda+\gamma > 1$, $\gamma > -1$ [2], [18]. For these polynomials the recursion coefficients are given by

$$a_n(d\alpha) = \frac{1}{2}\sqrt{(n+\gamma)(n+2\lambda+\gamma-1)} = \frac{n}{2}\left[1 + O(n^{-1})\right],$$

and thus

$$X_n(d\alpha)/n = 1 - 6^{-1/3} i_1 n^{-2/3} + o(n^{-2/3}). \tag{16}$$

For the case $\gamma = 0$, that is for

$$d\alpha(x) = |\Gamma(\lambda+ix)|^2 dx, \qquad -\infty < x < \infty, \tag{17}$$

Freud [5] gave an estimate for $X_n(d\alpha)$ that is weaker than (16). The measure associated with the symmetric Meixner polynomials [2] is closely related to (16).

The Associated Hermite Polynomials are orthogonal with respect to

$$d\alpha(x) = \exp(-x^2)\left|\gamma \int_0^\infty t^{\gamma-1} \exp(-2ixt - t^2)\, dt\right|^{-2} dx, \qquad -\infty < x < \infty,$$

where $\gamma \geq 0$ [1], and their recursion coefficients satisfy

$$a_n(d\alpha) = \sqrt{\frac{n+\gamma}{2}},$$

so that

$$X_n(d\alpha)n^{-1/2} = \sqrt{2} - 2^{-1/2}3^{-1/3} i_1 n^{-2/3} + o(n^{-2/3}). \tag{18}$$

When $\gamma = 0$ these polynomials are the Hermite polynomials.

The Hermite–Sonine Polynomials are orthogonal with respect to

$$d\alpha(x) = |x|^\lambda \exp(-x^2)\, dx, \qquad -\infty < x < \infty, \tag{19}$$

$\lambda > -1$ [2], [20] and in this case

$$a_n(d\alpha) = \sqrt{\frac{n+\theta_n}{2}} \tag{20}$$

where $\theta_{2m} = 0$, $\theta_{2m+1} = \lambda$. Thus $X_n(d\alpha)$ again satisfies (18).

The Freud polynomials are orthogonal with respect to

$$d\alpha(x)=|x|^{\lambda}\exp\left(-|x|^{m}\right)dx, \qquad -\infty<x<\infty,$$

where $\lambda>-1$ and $m>0$.

CONJECTURE.

$$a_n(d\alpha)=\frac{1}{2}\left[\sqrt{\pi}\,\frac{\Gamma(m/2)}{\Gamma((m+1)/2)}\right]^{1/m} n^{1/m}\left[1+O(n^{-1})\right]. \tag{21}$$

In view of (19) and (20) this holds for $m=2$. For $m=4$, $\lambda>-1$ this was proved by Lew–Quarles [6]. As mentioned before (21) also holds for $m=6$, $\lambda=0$ [9] and $m=$ even integer, $\lambda=0$ [10]. Thus

$$X_n(d\alpha)n^{-1/m}=\left[\sqrt{\pi}\,\frac{\Gamma(m/2)}{\Gamma((m+1)/2)}\right]^{1/m}\left(1-6^{-1/3}m^{-2/3}i_1n^{-2/3}+o(n^{-2/3})\right)$$

for $m=2, 4$, $\lambda>-1$ and m even, $\lambda=0$.

Finally, we point out that polynomials that are orthogonal on a half infinite line such as the Laguerre polynomials may be transformed by a quadratic change of the variable into symmetric orthogonal polynomials on the real line, and thus our theorem can be used to determine the size of their greatest zeros as well. Further recent results on the behavior of zeros of orthogonal polynomials are given in [3], [7], [11]–[16], [19] and [21].

One of the referees of this paper suggested finding explicit expression for the error term in (8) or at least an expression for an appropriate bound of the error term in (8). In [17, p. 408] F. W. J. Olver does provide such error analysis for the zeros of the Hermite polynomials. Naturally such a result would help to assess the computational feasibility of the approximation. At this time, however, we are unable to produce nontrivial error bounds in (8). Nonetheless we expect to return to this problem in a subsequent paper. We thank all the referees for their valuable remarks.

REFERENCES

[1] R. ASKEY AND J. WIMP, *Associated Laguerre and Hermite polynomials*, manuscript.

[2] T. S. CHIHARA, *An Introduction to Orthogonal Polynomials*, Gordon and Breach, New York, 1978.

[3] G. FREUD, *On the greatest zero of an orthogonal polynomial*, I, II, Acta Sci., Math. Szeged, 34 (1973), pp. 91–97 and 36 (1974), pp. 49–54.

[4] _____, *On the coefficients in the recursion formulae of orthogonal polynomials*, Proc. Royal Irish Acad. Sci., 76/A (1976), pp. 1–6.

[5] _____, *On the greatest zero of an orthogonal polynomial*, J. Approx. Theory, to appear.

[6] J. S. LEW AND D. A. QUARLES, *Nonnegative solutions of a nonlinear recurrence*, J. Approx. Theory, 38 (1983), pp. 357–379.

[7] D. S. LUBINSKY AND A.SHARIF, *On the largest zeros of orthogonal polynomials for certain weights*, Math. Comput., 41 (1983), pp. 199–202.

[8] AL. MAGNUS, *A proof of Freud's conjecture about the orthogonal polynomials related to* $|x|^{\rho}\exp(-x^{2m})$ *for integer* m, in Orthogonal Polynomials and Their Applications, C. Brezinski et al., eds., Lecture Notes in Mathematics, Springer, Berlin, 1985.

[9] A. MÁTÉ AND P. NEVAI, *Asymptotics for solutions of smooth recurrence equations*, Proc. Amer. Math. Soc., 93 (1985), pp. 423–429.

[10] A. MÁTÉ, P. NEVAI AND T. ZASLAVSKY, *Asymptotic expansions of ratios of coefficients of orthogonal polynomials with exponential weights*, Trans. Amer. Math. Soc., 287 (1985), pp. 495–505.

[11] H. N. MHASKAR AND E. B. SAFF, *Extremal problems for polynomials with exponential weights*, Trans. Amer. Math. Soc., 285 (1984), pp. 203–234.

[12] P. NEVAI, *Orthogonal polynomials on the real line associated with the weight* $|x|^{\alpha}\exp(-|x|^{\beta})$, I, Acta Math. Acad. Sci. Hungar., 24 (1973), pp. 335–342. (In Russian.)

[13] _____, *Distribution of zeros of orthogonal polynomials*, Trans. Amer. Math. Soc., (1979), pp. 341–361.

[14] _____, *Orthogonal polynomials*, Memoirs Amer. Math. Soc., Vol. 213, 1979.

[15] _____, *Orthogonal polynomials associated with* $\exp(-x^4)$, Second Edmonton Conference on Approximation Theory, CMS Conf. Proc. 3 (1983), pp. 263–285.

[16] P. NEVAI AND J. S. DEHESA, *On asymptotic average properties of zeros of orthogonal polynomials*, this Journal, 10 (1979), pp. 1184–1192.

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

[18] F. POLLACZEK, *Sur une famille de polynômes orthogonaux à quatre parametres*, C. R. Acad. Sci. Paris, 230 (1950), pp. 2254–2256.

[19] E. A. RAHMANOV, *On asymptotic properties of polynomials orthogonal on the real axis*, Math. USSR Sbo., 47 (1984), pp. 155–193.

[20] G. SZEGÖ, *Orthogonal Polynomials*, American Mathematical Society, Providence, RI, 1975.

[21] J. ULLMAN, *Orthogonal polynomials associated with an infinite interval*, Michigan Math. J., 27 (1980), pp. 353–363.

SIAM J. MATH. ANAL.
Vol. 17, No. 3, May 1986

016

ORTHOGONAL POLYNOMIALS, MEASURES AND RECURRENCE RELATIONS*

JOANNE DOMBROWSKI† AND PAUL NEVAI‡

Abstract. Properties of measures associated with orthogonal polynomials are investigated in terms of the coefficients of the three term recurrence formula satisfied by the orthogonal polynomials.

AMS(MOS) subject classification. Primary 43C05

Key words. orthogonal polynomials, recurrence relations, Szegö's theory

1. Introduction. Let $d\alpha$ be a positive measure on the real line with finite moments and infinite support, and let $\{p_n\}_{n=0}^{\infty}$, $p_n(x)=\gamma_n x^n+\cdots$, $\gamma_n>0$, be the system of orthonormal polynomials associated with $d\alpha$. The polynomials p_n satisfy the recurrence formula

$$xp_n=a_{n+1}p_{n+1}+b_n p_n+a_n p_{n-1}, \qquad n=0,1,\cdots, \tag{1}$$

where $p_{-1}=0$, $p_0=\gamma_0$, $a_0=0$, $a_n=\gamma_{n-1}/\gamma_n$ and

$$b_n=\int_{-\infty}^{\infty} xp_n^2(x)\,d\alpha(x).$$

By J. Favard's theorem [10, p. 60] every system of polynomials generated by (1) where $a_n>0$ $(n=1,2,\cdots)$ and $b_n\in\mathbb{R}$ is in fact a system of orthonormal polynomials. The corresponding measure $d\alpha$ is uniquely determined if and only if the associated moment problem has a unique solution, and the latter holds if, say, both sequences $\{a_n\}$ and $\{b_n\}$ are bounded. Recently there has been an upsurge in research activity concerning the determination of the relationship between orthogonal polynomials, recurrence relations and measures. Several such papers are listed in the references. In particular, R. Askey and M. Ismail [1, p. 102] asked whether it is true that if

$$a_n=\frac{1}{2}+\frac{c}{n}+O(n^{-2}) \quad \text{and} \quad b_n=0 \tag{2}$$

where $c>0$ then the absolutely continuous portion of the corresponding measure $d\alpha$ is in Szegö's class which means that $\log\alpha'(\cos t)\in L^1$. One of the main goals of this paper is to show that the Askey–Ismail problem can essentially be solved. More precisely, it follows from Theorem 3 below that if (2) is replaced by

$$a_n=\frac{1}{2}+\frac{c}{n}+\frac{d}{n^2}+o(n^{-2}) \quad \text{and} \quad b_n=0$$

where $c>0$ and $d\in\mathbb{R}$ then $\log\alpha'(\cos t)\in L^1$. While we suspect that condition (2) fails to imply the integrability of $\log\alpha'(\cos t)$, we do not have evidence supporting our claim at the present time. Let us point out that Theorem 2 in fact yields $\alpha'(x)\geqq \text{const}\sqrt{1-x^2}$, $|x|<1$, if only $a_n\downarrow 1/2$ and $b_n=0$. The latter is quite a surprise if compared with J. Shohat's result [26, p. 50] claiming that if $\operatorname{supp}(d\alpha)=[-1,1]$ then $\log\alpha'(\cos t)\in L^1$ if

* Received by the editors May 8, 1984.

† Department of Mathematics and Statistics, Wright State University, Dayton, Ohio 45435. This author's research was supported by the National Science Foundation under grant MCS-83-00882.

‡ Department of Mathematics, Ohio State University, Columbus, Ohio 43210.

and only if

$$\sum_{k=1}^{\infty}\left(a_k-\frac{1}{2}\right)<\infty, \qquad \sum_{k=1}^{\infty} b_k<\infty$$

and

$$\sum_{k=1}^{\infty}\left\{\left(a_k-\frac{1}{2}\right)^2+b_k^2\right\}<\infty.$$

The natural way to connect the recursion coefficients with the measure is via Stieltjes transforms (see e.g. [1]). However, it seems that this approach is feasible only when the recursion coefficients are given in terms of explicitly defined expressions such as rational functions of n. The general cases can better be handled by techniques introduced on one side in [5]–[9] and on the other side in [15], [16], [18], [19] and [20].

If p_n is generated by (1), then we define S_n by

$$S_n(x)=\sum_{k=0}^{n}\left\{\left[a_{k+1}^2-a_k^2\right]p_k^2(x)+a_k\left[b_k-b_{k-1}\right]p_{k-1}(x)p_k(x)\right\}. \tag{3}$$

All of our results are based on the formula

$$S_n(x)=a_{n+1}^2\left[p_n^2(x)-\frac{x-b_n}{a_{n+1}}p_n(x)p_{n+1}(x)+p_{n+1}^2(x)\right] \tag{4}$$

proved in [8]. (Caution: the notation in [8] is somewhat different!) The second of us believes that the significance of (4) cannot be overestimated, and it will play a fundamental role in future research on general orthogonal polynomials (see e.g. [22]). The other ingredient of this paper comes from [16] where the necessary spectral analysis was accomplished.

In order not to interrupt our forthcoming discussion, we first prove the following technical proposition. In what follows a_+ denotes the positive part of a and $\log^+$ and $\log^-$ are also defined in the usual way.

LEMMA 1. *Let* $\{a_n\}$ *and* $\{b_n\}$ *satisfy* $a_{n+1}\geqq\frac{1}{2}(1+|b_n|)$ *for* $n>N$ *and let* p_n *and* S_n *be defined by* (1) *and* (3) *respectively. Then*

$$(1-x^2)p_{n+1}^2(x)\leqq 4S_n(x), \qquad |x|\leqq 1, \tag{5}$$

$$(1-x^2)p_n^2(x)\leqq 4S_n(x), \qquad |x|\leqq 1, \tag{6}$$

$$\max_{|x|\leqq 1}p_{n+1}^2(x)\leqq 4(n+2)^2\max_{|x|\leqq 1}|S_n(x)|, \tag{7}$$

$$\max_{|x|\leqq 1}p_n^2(x)\leqq 4(n+1)^2\max_{|x|\leqq 1}|S_n(x)|, \tag{8}$$

$$0\leqq S_{n+1}(x)\leqq S_n(x)\exp\left\{4\frac{\left[a_{n+2}^2-a_{n+1}^2\right]_+ +a_{n+1}|b_{n+1}-b_n|}{1-x^2}\right\}, \qquad |x|\leqq 1, \tag{9}$$

and

$$\max_{|x|\leqq 1}S_{n+1}(x)\max_{|x|\leqq 1}S_n(x)\cdot\exp\left\{4(n+2)^2\left(\left[a_{n+2}^2-a_{n+1}^2\right]_+-a_{n+1}|b_{n+1}-b_n|\right)\right\} \tag{10}$$

hold for $n>N$.

Proof. By (4)

$$S_n(x)=a_{n+1}^2\left[p_n(x)-\frac{x-b_n}{2a_{n+1}}p_{n+1}(x)\right]^2+\frac{1}{4}\left[4a_{n+1}^2-(x-b_n)^2\right]p_{n+1}^2(x)$$

and

$$S_n(x)=a_{n+1}^2\left[p_{n+1}(x)-\frac{x-b_n}{2a_{n+1}}p_n(x)\right]^2+\frac{1}{4}\left[4a_{n+1}^2-(x-b_n)^2\right]p_n^2(x).$$

If $2a_{n+1}\geqq 1+|b_n|$ then $4a_{n+1}^2-(x-b_n)^2\geqq 1-x^2$ for $|x|\leqq 1$. Thus (5) and (6) are satisfied. Inequalities (7) and (8) follow from (5), (6) and Bernstein's theorem [17, p. 139]. Writing

$$S_{n+1}=S_n\left[a_{n+2}^2-a_{n+1}^2\right]p_{n+1}^2+a_{n+1}[b_{n+1}-b_n]p_np_{n+1}$$

and applying (5)–(8), inequalities (9) and (10) follow immediately.

THEOREM 1. *If* $\lim_{n\to\infty}=\frac{1}{2}$, $\lim_{n\to\infty}b_n=0$ *and*

$$\sum_{n=0}^{\infty}\{|a_{n+1}-a_n|+|b_{n+1}-b_n|\}<\infty \tag{11}$$

then the orthogonal polynomials p_n *generated by* (1) *and the corresponding measure* $d\alpha$ *satisfy*

(12)

$$\sum_{k=0}^{\infty}\{[a_{k+1}^2-a_k^2]p_k^2(x)+a_k[b_k-b_{k-1}]p_{k-1}(x)p_k(x)\}=\frac{\sqrt{1-x^2}}{2\pi\alpha'(x)},\qquad -1<x<1,$$

and the convergence is uniform on every closed subinterval of $(-1,1)$.

Proof. Theorem 1 follows immediately from (3), (4) and

$$\lim_{n\to\infty}\left[p_n^2(x)-\frac{x-b_n}{a_{n+1}}p_n(x)p_{n+1}(x)+p_{n+1}^2(x)\right]=\frac{2\sqrt{1-x^2}}{\pi\alpha'(x)}$$

which holds uniformly on every closed subinterval of $(-1,1)$ if (11) is satisfied [16].

THEOREM 2. *Let* $\{a_n\}_{n=1}^{\infty}$ *and* $\{b_n\}_{n=0}^{\infty}$ *satisfy* $a_{n+1}\geqq\frac{1}{2}(1+|b_n|)$ *for* $n>N$, $\lim_{n\to\infty}a_n=\frac{1}{2}$, $\lim_{n\to\infty}b_n=0$ *and*

$$\sum_{n=1}^{\infty}n^2\{[a_{n+1}-a_n]_+ +|b_{n+1}-b_n|\}<\infty.$$

Then there exist a constant $K>0$ *such that for the orthogonal polynomials* p_n *defined by* (1) *and for the associated measure* $d\alpha$ *we have*

$$\sqrt{1-x^2}\,|p_n(x)|\leqq K,\qquad -1\leqq x\leqq 1, \tag{13}$$

$n=1,2,\cdots$, *and*

$$\alpha'(x)\geqq K^{-1}\sqrt{1-x^2},\qquad -1\leqq x\leqq 1. \tag{14}$$

Proof. Repeated application of (10) shows that the sequence $\{S_n\}$ is uniformly bounded in $[-1,1]$ and then (13) follows from (6) whereas (14) follows from Theorem 1, (3) and (4).

Remark 1. The sharpness of Theorem 2 may best be illustrated by the ultraspherical polynomials which are orthogonal with respect to $d\alpha(x)=(1-x^2)^{\varepsilon}dx$ in $[-1,1]$. For these polynomials

$$a_n=\frac{1}{2}+\frac{1-4\varepsilon^2}{16}\frac{1}{n^2}+\frac{\text{const}}{n^3}+O(n^{-4})\quad\text{and}\quad b_n=0$$

so that the conditions of Theorem 2 are satisfied if and only if $|\varepsilon| \leqq \frac{1}{2}$ whereas (14) holds if and only if $\varepsilon \leqq \frac{1}{2}$.

THEOREM 3. *Let* $\{a_n\}_{n=1}^{\infty}$ *and* $\{b_n\}_{n=0}^{\infty}$ *satisfy* $a_{n+1} \geqq \frac{1}{2}(1+|b_n|)$ *for* $n > N$, $\lim_{n\to\infty} a_n = \frac{1}{2}$, $\lim_{n\to\infty} b_n = 0$ *and*

$$\sum_{k=1}^{\infty} n\{[a_{n+1}-a_n]_+ + |b_{n+1}-b_n|\} < \infty.$$

Let $d\alpha$ *be the measure associated with the orthogonal polynomials defined by* (1). *Then*

$$d\alpha(x) = w(x)\,dx + \textit{mass points outside } (-1,1) \tag{15}$$

where w *is positive and continuous in* $(-1,1)$, w *vanishes outside* $[-1,1]$ *and* w *belongs to Szegö's class, that is*

$$\int_0^{\pi} |\log w(\cos t)|\,dt < \infty. \tag{16}$$

Remark 2. If $\Sigma n\{|a_n - \frac{1}{2}| + |b_n|\} < \infty$ then the number of mass points in $d\alpha$ is finite and all such mass points are located outside $[-1,1]$ (see [4] and [11]).

Proof of Theorem 3. By the conditions $\Sigma(|a_{n+1}-a_n| + |b_{n+1}-b_n|) < \infty$ holds as well so that by the Theorem in [16] and by Blumenthal's result (see e.g. [18, Thm. 3.3.7, p. 23]) formula (15) holds with $w(>0) \in C(-1,1)$ and $\operatorname{supp}(w) = [-1,1]$. Therefore only (16) needs to be proved. Let δ_n be defined by

$$\delta_n = 4\left(\left[a_{n+2}^2 - a_{n+1}^2\right]_+ + a_{n+1}|b_{n+1}-b_n|\right).$$

Then by the assumptions made

$$\sum_{1}^{\infty} n\delta_n < \infty, \tag{17}$$

and applying (9) and (10) with $n > N$, we obtain

$$\begin{aligned}
&\int_0^{1-(n+1)^2} \frac{\log^+ S_{n+1}(x)}{\sqrt{1-x}}\,dx \\
&\quad \leqq \int_0^{1-n^{-2}} \frac{\log^+ S_n(x)}{\sqrt{1-x}}\,dx + \delta_n \int_1^{1-n^{-2}} \frac{dx}{(1-x)^{3/2}} \\
&\qquad + \log^+\left\{\max_{|x|\leqq 1} |S_{n+1}(x)|\right\} \int_{1-n^{-2}}^{1-(n+1)^{-2}} \frac{dx}{\sqrt{1-x}} \\
&\quad = \int_0^{1-n^{-2}} \frac{\log^+ S_n(s)}{\sqrt{1-x}}\,dx \\
&\qquad + 2(n-1)\delta_n + 2\left(\frac{1}{n}\right)\frac{1}{n+1} \log^+\left\{\max_{|x|\leqq 1} S_{n+1}(x)\right\} \\
&\quad \leqq \int_0^{1-n^{-2}} \frac{\log^+ S_n(x)}{\sqrt{1-x}}\,dx + 2(n-1)\delta_n + \frac{2}{n} \log^+\left\{\max_{|x|\leqq 1} S_n(x)\right\} \\
&\qquad - \frac{2}{n+1} \log^+\left\{\max_{|x|\leqq 1} S_{n+1}(x)\right\} + \frac{2(n+2)^2}{n}\delta_n.
\end{aligned}$$

Therefore

$$\int_0^{1-(n+1)^{-2}} \frac{\log^+ S_{n+1}(x)}{\sqrt{1-x}}\,dx \leqq \int_0^{1-n^{-2}} \frac{\log^+ S_n(x)}{\sqrt{1-x}}\,dx + 18n\delta_n + \frac{2}{n}\log^+\left\{\max_{|x|\leqq 1} S_n(x)\right\} - \frac{2}{n+1}\log^+\left\{\max_{|x|\leqq 1} S_{n+1}(x)\right\}$$

from which

$$\int_0^{1-(n+1)^{2}} \frac{\log^+ S_{n+1}(x)}{\sqrt{1-x}}\,dx \leqq \int_0^{1-(N+1)^{2}} \frac{\log^+ S_{n+1}(x)}{\sqrt{1-x}}\,dx + 18\sum_{k=N+1}^{n} k\delta_k + \frac{2}{N+1}\log^+\left\{\max_{|x|\leqq 1} S_{N+1}(x)\right\}$$

follows. Now letting $n\to\infty$ and applying (3), (17), Theorem 1 and Fatou's lemma, we see that

$$\int_0^{\pi/2} \log^- w(\cos t)\,dt > -\infty.$$

By similar arguments

$$\int_{\pi/2}^{\pi} \log^- w(\cos t)\,dt > -\infty$$

holds as well. By Jensen's inequality

$$\int_0^{\pi} \log w(\cos t)\,dt < \infty$$

and thus Theorem 3 has completely been proved.

2. Applications.

A. Ya. L. Geronimus [12]–[14] raised and solved the following problem. Let $m\geqq 0$ be a fixed integer and let $\{b_k\}_{k=0}^{\infty}$ and $\{a_k\}_{k=1}^{\infty}$ $(a_k>0)$ be given sequences such that $b_k=0$ for $k\geqq m$ and $a_k=\frac{1}{2}$ for $k>m$. Let $\{p_n\}_{n=0}^{\infty}$ be the orthogonal polynomial system generated by (1) and let $d\alpha$ be the corresponding measure. The problem is to find $d\alpha$. We will show that on the basis of our results $d\alpha$ can easily be found. It follows from (3) and (4) that

$$4S_n = p_n^2 - 2xp_n p_{n+1} + p_{n+1}^2, \qquad n\geqq m, \tag{18}$$

and

$$S_n = S_m, \qquad n\geqq m. \tag{19}$$

Thus by Theorem 1 and 3 and by Remark 2,

$$d\alpha(x) = \frac{2}{\pi}\frac{\sqrt{1-x^2}}{p_m^2(x) - 2xp_m(x)p_{m+1}(x) + p_{m+1}^2(x)}\chi(x) + \sum_{l=1}^{M} J_l\delta(x-x_l) \tag{20}$$

where χ is the characteristic function of $[-1,1]$ and x_l's are the mass points with mass $J_l>0$. It is well known [27] that for an arbitary system of orthogonal polynomials if the associated moment problem has a unique solution then x is a mass point for $d\alpha$ with

mass J if and only if $\Sigma p_k^2(x)<\infty$ and then

(21)

$$\frac{1}{J}=\sum_{k=0}^{\infty}p_k^2(x)=\lim_{k\to\infty}a_{k+1}\left[p'_{k+1}(x)p_k(x)-p_{k+1}(x)p'_k(x)\right]$$

$$=\lim_{k\to\infty}a_{k+1}\left[p_{k+1}(x)/p_k(x)\right]'p_k^2(x)=-\lim_{k\to\infty}a_{k+1}\left[p_k(x)/p_{k+1}(x)\right]'p_{k+1}^2(x).$$

Therefore by (18) and (19) the mass points x_l are zeros of S_m and hence $M\leqq 2m+2$. However not all zeros of S_m are in fact mass points. Applying the recurrence formula (1) and (3)–(4) we obtain

$$4S_m=p_{n+1}^2-p_np_{n+2},\qquad n\geqq m \tag{22}$$

so that

$$\frac{p_{n+1}}{p_n}=\frac{p_{m+1}}{p_m},\quad n\geqq m,\quad \text{if } S_m=0. \tag{23}$$

On the other hand, by (18)

$$\frac{p_{m+1}}{p_m}=x\pm\sqrt{x^2-1}\quad \text{if } S_m=0 \tag{24}$$

(here $\sqrt{x^2-1}>0$ if $x>1$ and $\sqrt{x^2-1}<0$ if $x<-1$). Thus by (21) x is a mass point if and only if $S_m(x)=0$ and $|p_{m+1}(x)|<|p_m(x)|$. By (22)

$$4S'_m=-\left(p_n/p_{n+1}\right)'p_{n+1}^2\frac{p_{n+2}}{p_{n+1}}-\left(p_{n+2}/p_{n+1}\right)'p_{n+1}^2\frac{p_n}{p_{n+1}},\qquad n\geqq m,$$

if $S_m=0$ so that by (21), (23) and (24)

$$2S'_m(x_l)=J_l^{-1}\left(x_l-\sqrt{x_l^2-1}\right)-J_l^{-1}\left(x_l+\sqrt{x_l^2-1}\right)=-2J_l^{-1}\sqrt{x_l^2-1}\,.$$

Thus the mass points x_l in (20) are those zeros of $p_m^2-2xp_mp_{m+1}+p_{m+1}^2$ for which $|p_{m+1}(x_l)|<|p_m(x_l)|$ and the corresponding mass J_l is given by

$$J_l=\frac{-\sqrt{x_l^2-1}}{S'_m(x_l)}=\left|\frac{\sqrt{x_l^2-1}}{S'_m(x_l)}\right|.$$

B. The previous analysis can be applied to the case when

$$\lim_{n\to\infty}\left|a_n-\frac{1}{2}\right|^{1/n}=0\quad\text{and}\quad\lim_{n\to\infty}|b_n|^{1/n}=0.$$

Without going into details we point out that in this case one can prove

$$\limsup_{n\to\infty}|p_n(x)|^{1/n}<\infty$$

uniformly on every compact set in the complex plane, and thus

$$S(x)=\lim_{n\to\infty}S_n(x)$$

is an entire function. The corresponding measure $d\alpha$ can be written as

$$d\alpha(x)=\frac{1}{2\pi}\frac{\sqrt{1-x^2}}{S(x)}\chi(x)\,dx+\sum_{x_l\in Z}\left|\frac{\sqrt{x_l^2-1}}{S'(x_l)}\right|\delta(x-x_l)$$

where χ is the characteristic function of $[-1,1]$ and Z is the collection of those zeros x_l of S for which $\lim_{n\to\infty}|p_{n+1}(x_l)/p_n(x_l)|<1$. Of course, Z is a finite set.

C. The Pollaczek polynomials satisfy (1) with

$$a_n^2=\frac{1}{4}\frac{(n+c)(n+2\lambda+c-1)}{(n+\lambda+a+c)(n+\lambda+a+c-1)},\qquad n=1,2,\cdots,$$

and

$$b_n=-\frac{b}{n+\lambda+a+c},\qquad n=0,1,2,\cdots,$$

where the parameters a, b, c and λ are chosen so that $b_n\in\mathbb{R}$ and $a_n>0$. Pollaczek [23] (see also [3]) investigated the case when either $a>|b|$, $2\lambda+c>0$, $c\geqq 0$ or $a>|b|$, $2\lambda+c\geqq 1$, $c>-1$ and determined that $d\alpha$ is absolutely continuous. Hence $d\alpha$ is completely described by (16). The explicit expression for α' [3, p. 185] shows that α' is not Szegö's class in this case. Since

$$a_n=\frac{1}{2}-\frac{a}{2n}+\frac{\text{const}}{n^2}+O(n^{-3})$$

holds, we see that the conditions of Theorem 2 are satisfied provided that $a<0$, $b=0$, and consequently

$$\alpha'(x)\geqq K^{-1}\sqrt{1-x^2},\qquad -1\leqq x\leqq 1,$$

with a suitably chosen positive constant K if $a<0$ and $b=0$. In particular, $\log\alpha'(\cos t)\in L^1$ in this case. Examples of Pollaczek polynomials with not necessarily absolutely continuous measures have been investigated in [1], [2], [24], and [29].

D. Let $\{a_n\}_{n=1}^{\infty}$ satisfy $0<a_n\leqq\frac{1}{2}$, $\lim_{n\to\infty}a_n=\frac{1}{2}$ and

$$\sum_{k=1}^{\infty}|a_{k+1}-a_k|<\infty,$$

and let $b_n=0$ for every n. Let $d\alpha$ be the measure associated with the orthogonal polynomials p_n which are defined by (1). It is well known that in this case $\operatorname{supp}(d\alpha)=[-1]$ (see e.g. [18, Thm. 3.3.7, p. 23]). Let us show that ± 1 are not mass points of $d\alpha$. If x is a mass point, then by (21) $\lim_{n\to\infty}|p_n(x)|=0$ so that there exists n_0 such that $|p_n(x)|\leqq p_{n_0}(x)$ for every n and $|p_n(x)|<|p_{n_0}(x)|$ for $n<n_0$. By the recurrence formula

$$|xp_{n_0}(x)|\leqq a_{n_0+1}|p_{n_0+1}(x)|+a_{n_0}|p_{n_0-1}(x)|<|p_{n_0}(x)|$$

and hence $|x|<1$, that is $x\neq\pm 1$. It has been shown in both [8] and [16] that $d\alpha$ is absolutely continuous in $(-1,1)$. Therefore $d\alpha$ is absolutely continuous on the whole real line and by Theorem 1

$$d\alpha(x)=\frac{1}{2\pi}\frac{\sqrt{1-x^2}}{\sum_{k=0}^{\infty}\left[a_{k+1}^2-a_k^2\right]p_k^2(x)}\chi(x)\,dx$$

where χ is the characteristic function of $[-1,1]$. By the previously quoted theorem of Shohat [26, p. 50] $\log \alpha'(\cos t) \in L^1$ if and only if $\Sigma(a_k - \frac{1}{2}) < \infty$.

E. If there exists N such that $a_n > a_{n+1}$ for $n > N$ and $\lim_{n\to\infty} a_n = \frac{1}{2}$ and if $b_n = 0$ for every n, then the conditions of Theorem 2 are satisfied. Hence $\alpha'(x) \geqq K^{-1}\sqrt{1-x^2}$, $|x| \leqq 1$, holds in this case.

REFERENCES

[1] R. ASKEY AND M. ISMAIL, *Recurrence Relations, Continued Fractions and Orthogonal Polynomials*, Memoirs of the Amer. Math. Soc., to appear.

[2] E. BANK AND M. ISMAIL, *The attractive Coulomb potential polynomials*, Constructive Approximation, to appear.

[3] T. S. CHIHARA, *An Introduction to Orthogonal Polynomials*, Gordon and Breach, New York, 1978.

[4] T. S. CHIHARA AND P. NEVAI, *Orthogonal polynomials and measures with finitely many point masses*, J. Aprox. Theory, 35 (1982), pp. 370–380.

[5] J. DOMBROWSKI, *Spectral properties of phase operators*, J. Math. Phys., 15 (1974), pp. 576–577.

[6] _____, *Spectral properties of real parts of weighted shift operators*, Indiana Univ. Math. J., 29 (1980), pp. 249–259.

[7] _____, *Tridiagonal matrix representations of cyclic self-adjoint operators*, Pacific J. Math., to appear.

[8] _____, *Tridiagonal matrix representations of cyclic self-adjoint operators* II, Pacific J. Math., to appear.

[9] J. DOMBROWSKI AND G. H. FRICKE, *The absolute continuity of phase operators*, Trans. Amer. Math. Soc., 213 (1975), pp. 363–372.

[10] G. FREUD, *Orthogonal Polynomials*, Pergamon Press, New York, 1971.

[11] J. S. GERONIMO AND K. M. CASE, *Scattering theory and polynomials orthogonal on the real line*, Trans. Amer. Math. Soc., 258 (1980), pp. 467–494.

[12] JA. L. GERONIMUS, *On some finite-difference equations and the corresponding systems of orthogonal polynomials*, Doklady Akad. Nauk SSSR, 29 (1940), pp. 536–538. (In French.)

[13] _____, *On some finite-difference equations and the corresponding systems of orthogonal polynomials*, Zap. Mat. Otd. Fiz.-Mat. Fak. i. Har'kov. Mat. Obsc. (4) 25 (1957), pp. 87–100. (In Russian.)

[14] JA. L. GERONIMUS AND G. SZEGÖ, *Two Papers on Special Functions*, Amer. Math. Soc. Translations, Series 2, Vol. 108, 1977.

[15] A. MÁTÉ AND P. NEVAI, *Sublinear perturbations of the differential equation* $y^{(n)} = 0$ *and of the analogous difference equation*, J. Differential Equations, to appear.

[16] _____, *Orthogonal polynomials and absolutely continuous measures*, in Approximation Theory, IV, C. K. Chui et al., eds., Academic Press, New York, 1983, pp. 611–617.

[17] I. P. NATANSON, *Constructive Function Theory*, Vol. 1, F. Ungar, New York, 1964.

[18] P. NEVAI, *Orthogonal Polynomials*, Memoirs of the Amer. Math. Soc. 213, 1979.

[19] _____, *Orthogonal polynomials defined by a recurrence relation*, Trans. Amer. Math. Soc., 250 (1979), pp. 369–384.

[20] _____, *Orthogonal polynomials defined by a recurrence relation*, Trans. Amer. Math. Soc., 250 (1979), pp. 369–384.

[21] _____, *Two of my favorite ways of obtaining asymptotics for orthogonal polynomials*, in Functional Analysis and Approximation, P. L. Butzer and B. Sz.-Nagy, eds., ISNM 65, Birkhauser Verlag, Basel, 1984, 417–436.

[22] _____, *Exact bounds for orthogonal polynomials*, J. Approx. Theory, 44 (1985), pp. 82–85.

[23] F. POLLACZEK, *On a four parameter family of orthogonal polynomials*, C. R. Acad. Sci. Paris, 230 (1950), pp. 2254–2256. (In French.)

[24] _____, *On a Generalization of Jacobi Polynomials*, Mémorial des Sciences Mathématiques, Vol. 121, Paris, 1956. (In French.)

[25] D. D. ROGERS, *Spectral properties of some tridiagonal matrices*, manuscript.

[26] J. A. SHOHAT, *General Theory of Chebyshev's Orthogonal Polynomials*, Mémorial des Sciences Mathématiques, Vol. 66, Paris, 1934. (In French.)

[27] J. A. SHOHAT AND J. D. TAMARKIN, *The Problem of Moments*, American Mathematical Society, Providence, RI, 1943.

[28] G. SZEGÖ, *Orthogonal Polynomials*, American Mathematical Society, Providence, RI, 1975.

[29] H. A. YAMANI AND W. P. REINHARDT, L^2 *discretization of the continuum: radical kinetic energy and Coulomb Hamiltonian*, Phys. Rev. A, 11 (1975), pp. 1144–1155.

SIAM J. MATH. ANAL.
Vol. 17, No. 4, July 1986

ON A CLASS OF SUPERLINEAR STURM–LIOUVILLE PROBLEMS WITH ARBITRARILY MANY SOLUTIONS*

BERNHARD RUF† AND SERGIO SOLIMINI‡

Abstract. We derive multiplicity results for autonomous superlinear ODE's of the form

$$-u''(x)=g(u(x))+t, \qquad x\in(0,\pi), \quad t\in R,$$
$$u(0)=u(\pi)=0,$$

with $g'(-\infty)<+\infty$ and $g'(+\infty)=+\infty$.

We show that for any given $n\in N$ there exist at least n solutions of the problem if t is sufficiently negative. The proof is carried out by using variational methods jointly with a rearrangement argument.

Key words. superlinear Sturm–Liouville problem, multiple solutions, variational problem, mountain pass solution, Steiner symmetrization

AMS(MOS) subject classifications. Primary 34B15; secondary 35A15

1. Introduction. We consider here the solvability of ordinary differential equations with superlinear nonlinearities of the following type:

$$-u''(x)=g(u(x))+t, \quad x\in(0,\pi), \qquad u(0)=u(\pi)=0, \tag{1}$$

with

$$\lim_{s\to+\infty}\frac{g(s)}{s}=+\infty \quad \text{and} \quad \limsup_{s\to-\infty}\frac{g(s)}{s}<1=\lambda_1,$$

i.e., the nonlinearity crosses *all* eigenvalues $\lambda_k=k^2$, $k\in N$, of the eigenvalue problem

$$-v''(x)=\lambda v(x), \quad x\in(0,\pi), \qquad v(0)=v(\pi)=0. \tag{2}$$

It is easy to see that (1) admits no solution if t is bigger than some t_0. On the other hand the results of E. N. Dancer [4] (see also [1], [5], [6]) ensure in particular that (1) admits at least two solutions for t smaller than t_0. We recall also the results of C. Scovel [12] for the equation

$$-u''=6u^2+t \quad \text{in } (0,\pi), \qquad u(0)=u(\pi)=0. \tag{3}$$

He has shown that for all $k\in N$ there exist values $t_k<\cdots<t_1$ such that for $t<t_k$ there exist k solutions of (3).

The aim of this paper is to prove such a result for equation (1) under general assumptions on g. The key idea to arrive at this result is the following: it is easy to see that (1) has a negative solution which is a local minimum of the associated functional. A second solution can then be found by the mountain pass theorem. We shall show that this mountain pass solution must change sign if t is large enough negative. A "rearrangement" of the minimizing paths then shows that the mountain pass solution

* Received by the editors August 14, 1984 and in revised form February 8, 1985.
† Forschungsinstitut für Mathematik, ETH, Zentrum, CH-8092 Zürich, Switzerland.
‡ International School for Advanced Studies, 33014 Trieste, Italy.

has in fact precisely one sign change. Finally, doing the same on the intervals $(0,\pi/n)$, we then show that these solutions can be joined to obtain solutions having $2n-1$ nodes.

We point out that it is essential for our method that equation (1) is autonomous (including that t is a constant). The nonautonomous equation as well as, of course, the corresponding partial differential equation, remain therefore open problems.

2. Statement of the result. We consider (1) under the following assumptions on g: $(g_1)\, g \in C^1(\mathbb{R})$ with

$$\limsup_{s \to -\infty} g'(s) =: g^- < 1, \qquad \lim_{s \to +\infty} g'(s) = +\infty,$$

$(g_2)\, g \in C^1(\mathbb{R})$ with

$$\limsup_{s \to -\infty} g'(s) < +\infty, \qquad \lim_{s \to +\infty} g'(s) = +\infty.$$

Of course (g_1) is a stronger assumption than (g_2) and it is the situation in which we are mainly interested. We shall assume the condition (g_1) throughout the paper until the proof of Theorem 1.

Our goal is to prove the following theorem:

THEOREM 1. *Assume* (g_2) *holds. Then for any* $k \in N$ *there exists* $t_k \in R$ *such that for* $t < t_k$ *problem* (1) *has at least* k *distinct solutions.*

Remark 2. In Theorem 1, when (g_1) holds, we actually prove that there exist a negative solution, and solutions with $1,3,5,\cdots,2k-1$ nodes. We shall also discuss the possible existence of positive solutions of (1).

We remark that if (g_1) holds, then there exists a constant t_0 such that for $t > t_0$ (1) has no solution. This follows from the following calculation:

Since $g(s) - g^- \cdot s \geqq -c$, $\forall s \in \mathbb{R}$, we have for $t > c$ and any solution u of (1)

$$-u'' - g^- u = (g - g^-)(u) + t > 0$$

and hence $u > 0$ by the maximum principle, since $g^- < 1$. Multiplying this equation by $\sin x$, we get

$$\begin{aligned}(1-g^-)(u,\sin x) &= ((g-g^-)(u),\sin x) + (t,\sin x) \\ &\geqq ((1-g^-)u,\sin x) - (d,\sin x) + (t,\sin x)\end{aligned}$$

which implies $d > t$. Hence there exists no solution for $t > d$.

On the other hand, if (g_1) does not hold, i.e. if $g^- < \lambda_1$, then the above estimate is not valid, and one then expects solutions for arbitrarily large t. In fact, for the following closely related equation

$$-u'' = \lambda u + (u^+)^p + t \sin x, \quad x \in (0,\pi), \qquad u(0) = u(\pi),$$

where $p > 1$, one has the following result (see B. Ruf and P. N. Srikanth [11]): If $\lambda \in (\lambda_k, \lambda_{k+1})$, then there exist at least $2k+2$ solutions for any $t > 0$.

It is to be expected that the same result holds for the inhomogeneity t instead of $t \sin x$ and for more general nonlinearities g with $g^- \in (\lambda_k, \lambda_{k+1})$, $\lim_{s \to +\infty}(g(s)/s) = +\infty$. □

We shall use variational methods on the space $E=H_0^1(0,\pi)$. We work with the functional I: $E\to\mathbb{R}$

$$I(u):=1/2\int_0^\pi |u'|^2-\int_0^\pi G(u)-t\int_0^\pi u,$$

where $G(s)=\int_0^s g(t)\,dt$ is the primitive of g.

It is clear that $I\in C^2(E,\mathbb{R})$, and we have

LEMMA 3. *The functional I satisfies the Palais–Smale condition, i.e. any sequence $(u_n)\subset E$ with $I(u_n)$ bounded and $I'(u_n)_{n\to\infty}\to 0$ in E', contains a convergent subsequence.*

Proof. Since $g^-<1$ we can find a constant $a\in(g^-,1)$ and a constant M such that $g(s)>as-M$, $\forall s<0$. Let (u_n) be a sequence satisfying the hypothesis, and set $r_n=-u_n''-g(u_n)-t$, which converges to zero in H^{-1}. Multiplying r_n by $u_n^-:=\max\{-u,0\}$ we get

$$\left|-\int_0^\pi u_n''u_n^- -\int_0^\pi (g(u_n)+t)u_n^-\right|\leqq C\|u_n^-\|_E$$

from which we get

$$\left|\int_{[u_n<0]}|u_n'|^2-\int_{[u_n<0]}(au_n^2-Mu_n)\right|\leqq C_1\|u_n^-\|_E$$

and hence

$$\|u_n^-\|_E^2-a\|u_n^-\|_E^2\leqq C_2\|u_n^-\|_E.$$

Therefore $\|u_n^-\|_E\leqq C$, $\forall n\in N$, and hence also $\forall\|u_n^-\|_{C^0}\leqq C$, $\forall n\in N$. This means that the u_n are bounded from below, and hence also $g(u_n)$ is bounded from below. Therefore there exists a $b>1$ (by (g_1)) such that

$$g(u_n)\geqq bu_n-M\quad\forall n\in N.$$

Now we multiply r_n by $\sin x$ and get the estimate

$$\int_0^\pi \sin x(g(u_n)-u_n)\,dx\leqq C.$$

Using the estimate for g we then obtain

$$\int_0^\pi \sin x(b-1)u_n\,dx\leqq C,$$

i.e.,

$$\int_0^\pi \sin x\cdot u_n\leqq C\quad\forall n\in N,$$

and then also $\int_0^\pi \sin x\cdot g(u_n)\leqq C$, $\forall n\in N$. Since $g(u_n)$ is bounded below, we finally get

$$\int_0^\pi \sin x|g(u_n)|\leqq C\quad\forall n\in N.$$

Let G be the Green's operator of $-\ddot{u}$ with the associated Green's function K. Then

$$u_n=G(g(u_n)+t+r_n)=\int_0^\pi K(x,y)(g(u_n(y))+t+r_n(y))\,dy.$$

It is easily seen that $K(x,y)$ satisfies the estimate

$$0 \leqq K(x,y) \leqq C \cdot \sin y.$$

Using this in the above equation, we obtain

$$u_n \leqq \int_0^\pi C \cdot \sin y \, |g(u_n(y))| + \int_0^\pi K(x,y)(t + r_n(y))\, dy \leqq \text{const} \quad \forall n \in N.$$

From this one concludes by standard arguments that (u_n) contains a convergent subsequence. □

We point out that we have actually proved that if $I'(u_n) \to 0$ in E', then (u_n) is precompact. Therefore we have in particular proved that the set of the solutions of (1), for a given t, is compact. This shows that Theorem 1 is in some sense optimal when (g_1) holds; in fact one cannot generically expect to get infinitely many solutions of (1) for the same value of t.

3. A priori estimates. We first recall how a negative solution of (1), which is a local minimum of I, can be found.

In fact, choosing $t < -g(0)$, we see that zero is a supersolution of (1). By J. Kazdan and F. W. Warner [7] we can fix a negative subsolution $\underline{u}$. Using arguments as in H. Hofer [6] and D. G. De Figueiredo and S. Solimini [5] one then proves the existence of a local minimum of I in $[\underline{u}, 0]$. We remark that from the subsequent arguments it will follow that any negative solution is a local minimum.

LEMMA 4. *For k, $\varepsilon \in R_+$ given, there exists $T(k,\varepsilon) \in R$ such that if $t < T(k,\varepsilon)$, then $u < -k$ on $[\varepsilon, \pi - \varepsilon]$ for any negative solution u of* (1).

Proof. Fix $k, \varepsilon \in R_+$. Now choose $t < 0$ such that

$$\sup_{[-k,0]} g(s) < -\frac{1}{2}t, \qquad (4k/-t)^{1/2} < \varepsilon.$$

We set

$$x_1 := \inf\left\{ x \in [0,\pi] \mid u(x) \leqq -k \text{ or } u(x) = \min_{s \in [0,\pi]} u(s) \right\}.$$

We estimate u on $[0, x_1]$. Note that in x_1 we have $u'(x_1) \leqq 0$ and $u(x_1) \geqq -k$, and that $\ddot{u} > -t/2$ in $[0, x_1]$. By the Taylor formula we therefore have

$$u(x) \geqq u(x_1) + u'(x_1)(x - x_1) - \frac{t}{4}(x - x_1)^2 \geqq -k - \frac{t}{4}(x - x_1)^2.$$

Setting $x = 0$, we get $0 \geqq -k - (t/4)x_1^2$, and hence $x_1 \leqq (4k/-t)^{1/2} < \varepsilon$.

Doing the same arguments on the other end of the interval, we find that if $x_2 = \sup\{ x \in [0,\pi] \mid u(x) \leqq -k \text{ or } u(x) = \min_{[0,\pi]} u \}$, then $x_2 \geqq \pi - \varepsilon$.

Finally we note that $u \leqq -k$ in the interval (x_1, x_2), because otherwise we get a local maximum $\bar{x}$ with $u(\bar{x}) \geqq -k$ which would imply $\ddot{u}(\bar{x}) > -\frac{1}{2}t > 0$: a contradiction.

PROPOSITION 5. *Assume (g_1). Then there exists a constant $\tau \in R$ such that for $t < \tau$ equation* (1) *has exactly one negative solution.*

Proof. We have only to show that for τ sufficiently negative there is at most one negative solution. Let us assume that there are two, say u_1 and u_2. Subtracting the two equations, we have

$$-(u_2 - u_1)'' = g(u_2) - g(u_1). \tag{4}$$

Set

$$a(x):=\begin{cases}\dfrac{g(u_2(x))-g(u_1(x))}{u_2(x)-u_1(x)} & \text{if } u_2(x)\neq u_1(x),\\ g'(u_1(x)) & \text{if } u_2(x)=u_1(x).\end{cases}$$

Now we choose $k\in N$ such that $s<-k$ implies $|(g'(s))^+-(g_-)^+|<\delta$, for some $\delta>0$. By Lemma 3 we now have for $t<T(k,\varepsilon)$

$$\int_0^\pi |(a(x))^+-(g^-)^+|<\delta\pi+2\varepsilon\cdot m,$$

where $m=\sup_{s\in R^-}|g'(s)|$.

We rewrite (4) as follows:

$$-(u_2-u_1)''=a(x)(u_2-u_1). \tag{5}$$

But the eigenvalues μ_i of

$$-v''=\mu a(x)v \tag{6}$$

are all strictly bigger than one, if we choose $\delta>0$ and $\varepsilon>0$ sufficiently small; this follows by A. Manes and A. M. Micheletti [9], since $(a(x))^+\to(g_-)^+<1$ in $L^1(0,\pi)$, and $(a(x))^+\geqq a(x)$. This shows that $u_2=u_1$. □

We now turn to the discussion of the existence of positive solutions. We shall show in the next proposition that (1) has no *strictly* positive solution for t large negative. This result is based on and generalizes a recent nonexistence result of M. Ramaswamy [10]. On the other hand, we shall show that there is a negatively diverging sequence of values of t for which (1) has positive solutions which have (degenerate) zeros in $(0,\pi)$. Finally, we remark that the autonomous character of the equation (1) is essential for the nonexistence result; in fact, we will construct a superlinearity for which (1), with t replaced by $t\sin x$, has a strictly positive solution for arbitrarily large negative t.

PROPOSITION 6. *There exists $t^*\in R$ such that* (1) *has no strictly positive solution for $t<t^*$.*

Proof. Let t be given, and let u be a strictly positive solution of (1). We set

$$\beta=\inf\{s\in R^+\mid -g(s)=t\},\qquad \alpha=\inf\{x\in(0,\pi)\mid u(x)=\beta\}.$$

Note that u is symmetric around $\pi/2$; this follows since by the unique solvability of the Cauchy problem every stationary point x of u is a symmetry point. Therefore the only stationary point of u is $\pi/2$, and u is strictly increasing in $(0,\pi/2)$.

We want to give upper estimates of α and $\pi/2-\alpha$.

First, we consider $v=u-u(\alpha)$ as a solution on $(\alpha,\pi-\alpha)$ of the equation

$$-v''=-u''=g(u)+t=\frac{g(u)-g(u(\alpha))}{u-u(\alpha)}v,\qquad v(\alpha)=v(\pi-\alpha)=0. \tag{7}$$

Since $v\geqq 0$ in $(\alpha,\pi-\alpha)$ it follows that the first eigenvalue of the coefficient $(g(u(x))-g(u(\alpha)))/(u(x)-u(\alpha))$, which we denote by $\mu_1((g(u)-g(u(\alpha)))/(u-u(\alpha)))$ (see e.g. [9]), is equal to one. But we have for all $x\in(\alpha,\pi-\alpha)$

$$\frac{g(u(x))-g(u(\alpha))}{u(x)-u(\alpha)}=g'(\xi)$$

for some $\xi > \mu(\alpha)$. Hence, if we let $g(u(\alpha)) = -t \to +\infty$, we get $u(\alpha) \to +\infty$ and hence $\xi \to +\infty$ which implies $|g'(\xi)| \to \infty$ by assumption. Therefore

$$(g(u)-g(u(\alpha)))/(u-u(\alpha)) \to +\infty$$

uniformly for $t \to -\infty$. This implies that the length of $[\alpha, \pi-\alpha]$ tends to zero as $t \to -\infty$ (see [9]).

To obtain an estimate for α, we consider $w(x) = (x/\alpha)u(\alpha) - u(x)$ on $[0,\alpha]$ as a solution of the equation

$$\begin{aligned} -w'' = u'' = -t - g(u) &= \frac{g(u(\alpha))-g(u)}{(x/\alpha)u(\alpha)-u} w, \\ w(0) = w(\alpha) &= 0. \end{aligned} \tag{8}$$

Since $u'' = -t - g(u) \geqq 0$, because $u < \beta$, u is convex in $[0,\alpha]$ and therefore $w > 0$, hence

$$\mu_1\left(\frac{g(u(\alpha))-g(u(x))}{(x/\alpha)u(\alpha)-u(x)} \right) = 1 \quad \text{on } [0,\alpha].$$

But, since $g(u(\alpha)) > g(u(x))$ by the choice of α,

$$\frac{g(u(\alpha))-g(u(x))}{(x/\alpha)u(\alpha)-u(x)} \geqq \frac{g(u(\alpha))-g(u(x))}{u(\alpha)-u(x)}.$$

Fixing $\delta := \inf\{s \in R^+ \mid g(s) \geqq -1/2t\}$ we get for $u(x) < \delta$:

$$\frac{g(u(\alpha))-g(u(x))}{u(\alpha)-u(x)} \geqq \frac{1}{2} \frac{g(u(\alpha))}{u(\alpha)}.$$

On the other hand, for $u(x) \geqq \delta$ we have

$$\frac{g(u(\alpha))-g(U(x))}{u(\alpha)-u(x)} = g'(\xi)$$

for some $\xi \geqq u(x) \geqq \delta$. Hence we get in any case that

$$\frac{g(u(\alpha))-g(u)}{(x/\alpha)u(\alpha)-u} \to \infty \quad \text{uniformly for } t \to -\infty,$$

since $t \to -\infty$ implies $g(u(\alpha))/u(\alpha) \to +\infty$ and $\delta \to +\infty$. Therefore we conclude that also $\alpha \to 0$ as $t \to -\infty$.

We have shown that for $t \to -\infty$ $\alpha \to 0$ and $\alpha \to \pi/2$. This contradiction shows that there cannot exist a positive solution for t large negative. □

Remark 7. The above proof works on any interval. If on the other hand a certain interval is given, the condition $\lim_{s \to +\infty} g'(x) = +\infty$ could be relaxed to a suitable finite lower estimate of this limit. Our proof in fact shows that with the assumption $g'(s)_{s \to +\infty} \to +\infty$ the length of an interval on which there exists a strictly positive solution tends to zero for $t \to -\infty$.

This remark leads to the observation that Proposition 5 is not true if we omit the requirement that u is a strictly positive solution. In fact one has:

PROPOSITION 8. *There exists a sequence* $t_{n(n \to \infty)} \to -\infty$ *such that* (1) *has a positive solution* (*with interior degenerate zeros*) *for* $t = t_n$.

Proof. For given t consider the Cauchy problem

$$-u'' = g(u) + t, \qquad u(0) = u'(0) = 0. \tag{9}$$

Multiplying this equation by u' and integrating, we get

$$-|u'|^2 = G(u) + tu, \tag{10}$$

where $G(s) = \int_0^s g(t)\,dt$ denotes the primitive of g. Also, since g is asymptotically monotone, we have the relation

$$G(s) < g(s)\cdot s \quad \text{for } s \text{ large.} \tag{11}$$

Let now $\bar{s} \in R^+$ be the smallest value such that $G(\bar{s}) = t\bar{s}$. Then, assuming $t < -g(0)$, the solution of (9) starts monotonically increasing in zero, and by (10) it grows monotonically until it either reaches the value $\bar{s}$, or it tends monotonically to some limit. However, this last case is not possible, since this limit has to be $\bar{s}$ (since $u'(t)$ must tend to zero for $t \to -\infty$) and then

$$-u''(x) = g(u(x)) + t \to g(\bar{s}) + t > 0,$$

for t large negative, because this implies $\bar{s}$ large, and hence $g(\bar{s}) > G(\bar{s})/\bar{s} = -t$. Such an estimate for u'' is clearly impossible.

Therefore u must reach $\bar{s}$ in some finite point, say $x(t)/2$, and since this value is then again a symmetry point, we have that u is a positive solution on $[0, x(t)]$.

Finally, since u is a strictly positive solution on $[0, x(t)]$, the Remark 7 implies that $x(t) \to 0$ for $t \to -\infty$, and one sees easily that $x(t)$ depends continuously on t. Therefore there exists a sequence $t_n \to -\infty$ such that $x(t_n) = \pi/n$, for n large enough, and hence we can join the solutions on $[0, \pi/n]$ to obtain positive solutions on $[0, \pi]$ having degenerate zeros in the points $j\pi/n$, $0 \leqq j \leqq n$. □

Remark 9. The autonomous character of (1) is essential for Proposition 8. In fact, the following construction leads to a convex superlinearity f satisfying (g_1) and such that for

$$-u''(x) = f(u(x)) + t\sin x, \qquad u(0) = u(\pi) = 0 \tag{12}$$

the proposition does not hold.

We construct a sequence of asymptotically positively homogeneous problems; we choose a strictly convex function f_0 with $\lambda_1 < f_0'(+\infty) < \lambda_2$. By A. C. Lazer and P. J. McKenna [8] and S. Solimini [13, §2] there exists $t_0 < 0$ such that the corresponding problem (12_0) has a strictly positive solution u_0. Setting $m_0 = \max_{x \in [0,\pi]} u_0(x)$, we choose a strictly convex function f_1 with

$$f_1(s) \quad \begin{cases} = f_0(s), & s < m_0, \\ \text{strictly convex with } \lambda_2 < f_0'(+\infty) < \lambda_3. \end{cases}$$

For the problem (12_1) there then exists $t_1 < -1$ such that there exists a strictly positive solution u_1, with $m_1 = \max_{x \in [0,\pi]} u_1(x)$. Repeating this argument, we can find a strictly convex function f_n, $\forall n \in N$, such that (12_n) has a strictly positive solution u_n for $t_n \leqq -n$, and u_n is also a solution for (12_m), for $m > n$, and for the problem (12) with $f = \sup_{n \in N}\{f_n\}$.

4. A solution with one sign change. Here we want to show that the mountain pass solution can be assumed to have exactly one sign change.

For given t, we choose $r \in R^+$ such that

$$I(r\cdot \sin x) \leqq I(\underline{u}),$$

where $\underline{u}$ is the negative solution of (1). Now let

$$\Gamma = \{\gamma \in C([0,1], E) \mid \gamma(0) = \underline{u}, \gamma(1) = r \sin x\}$$

and set

$$C = \inf_{\gamma \in \Gamma[0,1]} \max I(\gamma(t)) > I(\underline{u}). \tag{13}$$

By the mountain pass theorem as given in A. Ambrosetti and P. Rabinowitz [2] one has in particular the following result.

THEOREM. *Let* $(\gamma_n) \subset \Gamma$ *such that*

$$\max_{[0,1]} I(\gamma_n(t))_{n \to \infty} \to C.$$

Then there exists a subsequence (γ_{n_k}) *such that one can find a sequence* (u_{n_k}) *with* $u_{n_k} \in \gamma_{n_k}$ *and* $u_{n_k} \to u$, *where* u *is a critical point of* I *at level* C.

We introduce the following class of functions:

$$\mathscr{S} = \{u \in E \mid (u(x) > 0, u(y) < 0) \Rightarrow x < y,$$
$$(u(x) > 0, u(y) > 0,\ x < z < y) \Rightarrow u(z) \geqq \min\{u(x), u(y)\}\}.$$

Note that if we find a solution v in $\mathscr{S}$ at level C, then this solution has exactly one nodal point. In fact, v cannot be negative, since by Proposition 5 there exists only one negative solution of (1), and it is at a lower level. Moreover, for t large negative, v cannot be positive, since by Proposition 6 it has to have interior zeros which contradicts the second requirement of $\mathscr{S}$. Finally, the condition of $\mathscr{S}$ says that v can have at most one sign change.

Note that $\mathscr{S}$ is closed for the pointwise convergence (and therefore in E).

We now introduce a procedure which will assign to any $u \in E$ a set of "rearranged" functions lying in $\mathscr{S}$.

We denote by σ: $E \to H^1(R)$ the *Steiner symmetrization* of a positive function in E around zero, i.e. for $u \in E$, $u \geqq 0$, $\sigma(u)$ is the unique function which is even, nonincreasing on $[0, \infty)$ and such that

$$\text{meas}\{x \mid \sigma(u(x)) \geqq y\} = \text{meas}\{x \mid u(x) \geqq y\}, \qquad \forall y \in R,$$

where meas A stands for the Lebesgue measure of A. $\sigma(u)$ satisfies

(a) For h: $[0, \infty) \to R$ continuous: $\int_R h(\sigma(u))\,dx := \int_R h(u)\,dx$

(b) $\sigma(u) \in H^1(R)$,

$$\int_R \left|\frac{d\sigma(u)}{dx}\right|^2 dx \leqq \int_R \left|\frac{du}{dx}\right|^2 dx,$$

(c) σ is strongly continuous in the H^1-norm (see J. M. Coron [3]).

We introduce some more notation. For $u \in E$, let $\alpha_+ = \mu\{u > 0\}$ and $\alpha_- = \mu\{u < 0\}$. Then $\alpha_+ + \alpha_- \leqq \pi$, and we can define the following set

$$\mathscr{P}(u) = \Big\{(p_+, p_-) \in [0,\pi]^2 \mid d(p_\pm, \{0,\pi\})$$
$$\geqq \frac{1}{2}\alpha_\pm, d(p_+, p_-) \geqq \frac{1}{2}(\alpha_+ + \alpha_-),\ p_+ \leqq p_-\Big\}.$$

Clearly, $\mathscr{P}(u)$ is nonempty and convex.

For $u \in E$ and $(p_+, p_-) \in \mathscr{P}(u)$ we now define

$$s(u, p_+, p_-) = \begin{cases} \pm[\sigma(u^{\pm})](x - p_{\pm}) & \text{if } d(x, p_{\pm}) \leqq \frac{1}{2}\alpha_{\pm}, \\ 0 & \text{otherwise} \end{cases}$$

and we set

$$S(u) = \bigcup_{(p_+, p_-) \in \mathscr{P}(u)} \{s(u, p_+, p_-)\}.$$

By the properties of the Steiner symmetrization we have

$$I(v) \leqq I(u) \quad \forall v \in S(u).$$

PROPOSITION 10. *Let $(u_n) \subset E$ be a sequence with $u_n \to u$ in E, and let $(p_+^n, p_-^n) \in \mathscr{P}(u_n)$ with $p_{\pm}^n \underset{n \to \infty}{\to} p_{\pm}$. Then $(p_+, p_-) \in \mathscr{P}(u)$ and*

$$v_n = s(u_n, p_+^n, p_-^n) \to v = s(u, p_+, p_-) \quad \text{in } E.$$

Proof. Since $u \to u^{\pm}$, σ and the translation operator are continuous in E, the result follows easily.

COROLLARY 11. *For any $u \in E$ the set $S(u)$ is connected in E.*

Proof. The map $s(u, \cdot, \cdot)$: $\mathscr{P}(u) \to E$ is continuous by Proposition 10, and since $\mathscr{P}(u)$ is convex, we obtain that $S(u)$ is connected.

COROLLARY 12. *If $(u_n) \subset E$ is a sequence with $u_n \to u$ in E, and $v_n \in S(u_n)$, then there exists a subsequence (v_{n_k}) with $v_{n_k} \to v$ in E for some $v \in S(u)$.*

Proof. Since $v_n = s(u_n, p_+^n, p_-^n)$ for some $(p_+^n, p_-^n) \in \mathscr{P}(u_n)$ we can choose a convergent subsequence $(p_+^{n_k}, p_-^{n_k}) \to (p_+, p_-)$ and apply Proposition 10 to obtain

$$v_{n_k} \to v = s(u, p_+, p_-).$$

PROPOSITION 13. *If γ is connected in E, then also $\bigcup_{u \in \gamma} S(u)$ is connected.*

Proof. Take any two closed sets C_1, C_2 with $\bigcup_{u \in \gamma} S(u) = C_1 \cup C_2$ and $C_1 \cap C_2 \cap \bigcup_{u \in \Gamma} S(u) = \varnothing$. For $u \in \gamma$ we have $S(u) \in C_1$ or $S(u) \in C_2$, since $S(u)$ is connected. Now let $\gamma_i = \{u \in \gamma \mid S(u) \subset C_i\}$, $i = 1, 2$. The γ_i are relatively closed because if $(u_n) \subset \gamma_i$ with $u_n \to u$ then there exists $(v_n) \subset S(u_n) \cap C_i$ of which a subsequence (v_{n_k}) converges to $v \in S(u)$ by Proposition 10. But since C_i is closed we have $v \in C_i$ and hence $S(u) \subset C_i$. By construction we have $\gamma_1 \cup \gamma_2 = \gamma$ with $\gamma_1 \cap \gamma_2 = \varnothing$. Therefore $\gamma = \gamma_i$ with $i = 1$ or $i = 2$, since γ is connected, and hence $\bigcup_{\gamma} S(u) \subset C_i$, for $i = 1$ or 2.

PROPOSITION 14. *Fix $t < -f(0)$ and let C be the critical value given in* (13). *Then there exists a critical point u with $I(u) = C$ and $u \in \mathscr{S}$.*

Proof. By the definition of C we can pick a sequence $\gamma_n \subset \Gamma$ such that $\max_{t \in [0, \pi]} I(\gamma_n(t))_{n \to \infty} \to C$. We have seen that $\bigcup_{u \in \gamma_n} S(u)$ is connected and that $\sup I(\bigcup_{u \in \gamma_n} S(u)) \leqq \sup I(\gamma_n(t))$, $\forall n \in N$. We can find an open, connected neighborhood A_n of $\bigcup_{\gamma_n} S(u)$ such that $d(x, \bigcup_{\gamma_n} S(u)) \leqq 1/n$, $\forall x \in A_n$ and that $\sup I(A_n) \leqq \sup I(\gamma) + 1/n$. Therefore one finds a path $\tilde{\gamma}_n \subset A_n$ with $\tilde{\gamma}_n(0) = \underline{u}$ and $\tilde{\gamma}_n(1) = r \cdot \sin x$ and such that

$$\sup_{t \in [0, 1]} I(\tilde{\gamma}_n(t)) \leqq \sup_{t \in [0, 1]} I(\gamma_n(t)) + \frac{1}{n}, \quad \forall n \in N.$$

By the mountain pass theorem we can therefore find a sequence $u_{n_k} \in \tilde{\gamma}_{n_k}$ with $u_{n_k} \to u$, where u is again a critical point at level C. Since moreover $d(u_{n_k}, \mathscr{S}) \leqq 1/n$ and $\mathscr{S}$ is closed, we have $u \in \mathscr{S}$. □

By our previous remarks we have therefore found a solution u with exactly one sign change, provided t is sufficiently negative.

5. Joining of solutions defined on $(0,\pi/n)$. The purpose of proving the existence of solutions which have exactly one sign change is that such solutions (if they are defined on a suitable sub-interval, say $(0,\pi/n)$) can be joined to solutions having many nodes. This is based on the following observation.

LEMMA 15. *Let u be a solution of* (1), *and let $x\in(0,\pi]$ be a zero of u. Then* $|u'(x)|=|u'(0)|$.

Proof. Multiplying (1) by u' we have

$$-\frac{1}{2}\frac{d}{dx}|u'(x)|^2=\frac{d}{dx}(F(u(x))+tu(x)).$$

Therefore (choosing $F(0)=0$) by integration

$$\frac{1}{2}|u'(x)|^2=\frac{1}{2}|u'(0)|^2-F(u(x))-tu(x).$$

Hence, if $u(x)=0$: $|u'(x)|=|u'(0)|$. □

We are now in the position to complete the proof of Theorem 1.

Proof of Theorem 1. First assume (g_1) and let $k\in N$ be given. Consider (1) on the interval $(0,\pi/m)$, $m=1,\cdots,k-1$. Choosing t smaller than some T (depending on k), we find solutions u_m on $(0,\pi/m)$ which change sign exactly once (Proposition 14). By Lemma 15 we know that $u_m'(0)=u_m'(\pi/m)$. Therefore, u_m can be extended to $(0,\pi)$ by setting

$$\tilde{u}_m(x)=\begin{cases}u(x), & x\in(0,\pi/m),\\ u\left(x-\dfrac{\pi}{m}\right), & x\in(\pi/m,2\pi/m),\\ u\left(x-\dfrac{\pi(m-1)}{m}\right), & x\in\left(\dfrac{\pi(m-1)}{m},\pi\right).\end{cases}$$

Hence we obtain the negative solution on $(0,\pi)$ and $k-1$ solutions having $2m-1$ $(m=1,\cdots,k-1)$ sign changes, respectively.

If we replace (g_1) by (g_2) then the argument works for large m and therefore the statement of the theorem still holds. □

Acknowledgment. We should like to thank E. N. Dancer and O. Kavian for their kind help and stimulating discussions.

Note added in proof. A similar result has been obtained simultaneously by A. Castro and R. Shivaji, Multiple solutions for a Dirichlet problem with jumping nonlinearities, to appear in the Proceedings of Conference on Diff. Equations, Arlington, Texas, 1984.

REFERENCES

[1] H. AMANN AND P. HESS, *A multiplicity result for a class of elliptic boundary value problems*, Proc. Royal Soc. Edinburgh, 84A (1979), pp. 145–151.

[2] A. AMBROSETTI AND P. RABINOWITZ, *Dual variational methods in critical point theory and applications*, J. Funct. Anal., 14 (1973), pp. 349–381.

[3] J. M. CORON, *The continuity of the rearrangement in $W^{1,p}(R)$*, Ann. Sci. Nor. Serie IV, XI, 1, (1984), pp. 57–86.

[4] E. N. DANCER, *On the ranges of certain weakly nonlinear elliptic partial differential equations*, J. Math. Pures Appl., 57 (1978), pp. 351–366.

[5] D. DE FIGUEIREDO AND S. SOLIMINI, *A variational approach to superline elliptic problems*, Comm. Partial Differential Equations, to appear.

[6] H. HOFER, *Variational and topological methods in partially ordered Hilbert spaces*, Math. Ann., 261 (1981), pp. 493–514.

[7] J. KAZDAN AND F. W. WARNER, *Remarks on some quasilinear elliptic equations*, Comm. Pure Appl. Math., XXVIII (1975), pp. 567–597.

[8] A. LAZER AND P. J. MCKENNA, *On the number of solutions of a nonlinear Dirichlet problem*, J. Math. Anal. Appl., 84 (1981), pp. 282–294.

[9] A. MANES AND A. M. MICHELETTI, *Un' estensione della teoria variazionale classica degli autovalori per operatori ellittici del second ordine*, Boll. Un. Mat. Ital., (4) 7 (1973), pp. 285–301.

[10] M. RAMASWAMY, *Quelques problèmes non-linéaires: homogénéisation et comportement global des solutions d'une équation différentielle non linéaire*, Thesis—Université Pierre et Marie Curie—Paris VI.

[11] B. RUF AND P. N. SRIKANTH, *Multiplicity results for* ODE's *with nonlinearities crossing all but a finite number of eigenvalues*, to appear in Nonlinear Anal., TMA.

[12] C. SCOVEL, *Geometry of some nonlinear differential operators*, Ph.D. thesis, Courant Institute, New York Univ., New York.

[13] S. SOLIMINI, *Existence of a third solution for a class of b.v.p. with jumping nonlinearities*, Nonlin. Anal. TMA, 7 (1983), pp. 917–927.

SIAM J. MATH. ANAL.
Vol. 17, No. 4, July 1986

002

CHARACTERIZATIONS OF THE FRIEDRICHS EXTENSIONS OF SINGULAR STURM–LIOUVILLE EXPRESSIONS*

HANS G. KAPER†, MAN KAM KWONG†‡ AND ANTON ZETTL†‡

Abstract. A method is presented to characterize selfadjoint realizations of a singular Sturm–Liouville differential expression on a finite interval, where the singularities are of limit-circle type.

Key words. Sturm–Liouville differential operators, singularities of limit-circle type, selfadjoint realizations, Friedrichs extension

AMS(MOS) subject classifications. Primary 34B25, 47E05

1. Introduction. In this article we present a new method for defining selfadjoint realizations of a certain class of singular Sturm–Liouville differential expressions

$$\tau = -\frac{d}{dt}p(t)\frac{d}{dt} + q(t) \tag{1}$$

on a finite interval (a,b). We assume throughout that p and q are measurable and real-valued functions on (a,b) which satisfy the minimal conditions

$$p^{-1}, q \in L^1_{\text{loc}}(a,b). \tag{2}$$

Moreover, we assume that p is positive,

$$p(t) > 0 \quad \text{a.e. on } (a,b). \tag{3}$$

Thus, τ is a quasi-differential expression in the sense of Naimark [1, §V.1]. A function $\bar{y}$ is said to be a solution of the equation $\tau y = 0$ if (i) $\bar{y}$ is absolutely continuous on (a,b), (ii) $p\bar{y}'$ is equal a.e. on (a,b) to an absolutely continuous function (which, with a slight abuse of notation, we denote by the symbol $p\bar{y}'$), and (iii) the identity $-(p\bar{y}')'(t) + q(t)\bar{y}(t) = 0$ holds a.e. on (a,b).

The right endpoint b is said to be a regular endpoint for τ if

$$p^{-1}, q \in L^1(c,b) \quad \text{for some } c \in (a,b). \tag{4}$$

Similarly, the left endpoint a is regular if

$$p^{-1}, q \in L^1(a,c) \quad \text{for some } c \in (a,b). \tag{5}$$

If both endpoints are regular, then the differential expression τ is called regular; otherwise, it is called singular. Note that, for τ to be regular, neither p^{-1} nor q need to be bounded on (a,b).

All solutions $\bar{y}$ of a regular Sturm–Liouville equation $\tau y = 0$ are continuous on $[a,b]$, and the same property holds for the function $p\bar{y}'$. Hence, boundary values can be assigned to these functions. The characterization of those boundary conditions which

*Received by the editors August 17, 1984, and in revised form February 23, 1985. This work was supported by the Applied Mathematical Sciences Research Program (KC-04-02) of the Office of Energy Research of the U. S. Department of Energy under Contract W-31-109-ENG-38.

†Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, Illinois 60439.

‡Permanent address: Department of Mathematical Sciences, Northern Illinois University, DeKalb, Illinois 60115

give rise to selfadjoint realizations of a regular differential expression in the Hilbert space $L^2(a,b)$ is well known and can be found, for example, in the monographs by Akhiezer and Glazman [2, Appendix II] and Naimark [1, §5.18].

The study of singular differential expressions is considerably more difficult. The solutions of a singular Sturm–Liouville equation $\tau y = 0$ generally exhibit singularities near the endpoints, so one cannot assign boundary values there. Weyl [3] has developed a theory for the construction of selfadjoint realizations of singular differential expressions, which is based on a distinction between singularities of limit-circle type and those of limit-point type. The characterizations are, however, not concrete and therefore difficult to apply. The same remarks can be made for the theory developed by Titchmarsh [4].

In this article we present a new method for characterizing selfadjoint realizations of singular Sturm–Liouville differential expressions τ of the form (1), where q is bounded and the singularity at either endpoint is of limit-circle type. We limit the discussion to the case of one singular endpoint; the extension of the method to cases where both endpoints are singular is straightforward. Specifically, we assume that the coefficients p and q satisfy, in addition to (2), (3), and (4), the conditions

$$\int_t^b p^{-1}(s)\,ds = O\big((t-a)^{-\gamma}\big) \quad \text{as } t \downarrow a, \quad \gamma \in \left(0, \frac{1}{2}\right), \tag{6}$$

$$q \in L^\infty(a,b). \tag{7}$$

Thus, b is a regular endpoint and a is a singular endpoint for τ. For bounded potentials q, the condition (6) is both necessary and sufficient for the singularity at a to be of limit-circle type.

A selfadjoint realization of τ in $L^2(a,b)$ requires the specification of two boundary conditions, one at the regular endpoint b and one at the singular endpoint a. At b we impose a condition of the usual type,

$$B_1 y(b) + B_2(py')(b) = 0, \qquad B_1^2 + B_2^2 \neq 0. \tag{8}$$

Given such a condition, there are an infinite number of conditions at a which give rise to a selfadjoint realization of τ in $L^2(a,b)$. The particular condition

$$\lim_{t \downarrow a} y(t) \text{ exists and is finite} \tag{9}$$

is known to generate a selfadjoint realization which coincides with the Friedrichs extension of the minimal operator in $L^2(a,b)$ associated with τ. Although the condition (9) is often referred to as the "natural" one, relating it to the Weyl or Titchmarsh theory of singular Sturm–Liouville problems is nontrivial.

As we will demonstrate, (9) is but one of several equivalent characterizations of the same selfadjoint realization of τ. These characterizations follow in a systematic way from a particular representation of the elements in the domain of the maximal operator defined by τ. The procedure sheds some light on the role that the particular condition (9) plays within the general framework of Weyl's theory.

2. Characterization of the Friedrichs extension. Let $\phi : (a,b) \to \mathbb{R}$ be defined by the expression

$$\phi(t) = 1 + \int_t^b p^{-1}(s)\,ds, \qquad t \in (a,b). \tag{10}$$

Then $\phi \in L^2(a,b)$, because (a,b) is finite and p satisfies (6). Furthermore, $\phi(t) \geqq 1$ for all $t \in (a,b)$ and $\phi'(t) = -p^{-1}(t)$ a.e. on (a,b).

Let M be the maximal operator associated with τ,

$$My = \tau y, \qquad y \in \operatorname{dom} M, \tag{11}$$

where $\operatorname{dom} M = \{ y \in L^2(a,b) : y \text{ and } py' \text{ locally absolutely continuous on } (a,b) \text{ and } \tau y \in L^2(a,b)\}$. The following lemma gives a representation of the elements of $\operatorname{dom} M$.

LEMMA 1. *For every $y \in \operatorname{dom} M$ there exist two constants c and d and an element $g \in L^2(a,b)$, such that*

$$y(t) = c\phi(t) + d + \int_a^t (\phi(t) - \phi(s)) g(s)\, ds, \qquad t \in (a,b), \tag{12}$$

$$y'(t) = -cp^{-1}(t) - p^{-1}(t) \int_a^t g(s)\, ds, \qquad t \in (a,b). \tag{13}$$

Proof. Because q is bounded, $\operatorname{dom} M$ consists of those $y \in L^2(a,b)$ for which y and py' are locally absolutely continuous on (a,b) and $(py')' \in L^2(a,b)$. Hence, for every $y \in \operatorname{dom} M$ there exists a $g \in L^2(a,b)$ such that $-(py')' = g$. Integration of this identity gives the representations (12) and (13). □

Selfadjoint realizations T of τ are obtained by restricting M. The restrictions result in constraints on the element g and the constants c and d in the representation (12). The boundary condition (8) imposes one such constraint, viz.,

$$(B_1 - B_2)c + B_1 d = \int_a^b (B_1 - B_2 - B_1\phi(s)) g(s)\, ds. \tag{14}$$

Another constraint is obtained by imposing a "boundary condition" at the singular endpoint. For example, the condition (9) leads to the constraint $c = 0$. The following lemma explores the ramifications of this constraint.

LEMMA 2. *Let $y \in \operatorname{dom} M$. Then the following conditions are equivalent*:

(i) *y has a representation of the form* (12) *with* $c = 0$;
(ii) *y is bounded on (a,b)*;
(iii) $\lim_{t \downarrow a} y(t)$ *exists and is finite*;
(iv) $\lim_{t \downarrow a} (t-a)^{\gamma} y(t) = 0$;
(v) $\lim_{t \downarrow a} (py')(t) = 0$;
(vi) $\lim_{t \downarrow a} (t-a)^{-\alpha} (py')(t) = 0$ *for any* $\alpha \in (0, \frac{1}{2})$;
(vii) $p^{1/2} y' \in L^2(a,b)$;
(viii) $(t-a)^{-\alpha/2} p^{1/2} y' \in L^2(a,b)$ *for any* $\alpha \in (0, \frac{1}{2})$;
(ix) $y' \in L^1(a,b)$.

Proof. (i) $\Leftrightarrow$ (ii). Elementary estimates yield the inequalities

$$\left| \int_a^t (\phi(t) - \phi(s)) g(s)\, ds \right| \leqq \phi(t) \left| \int_a^t g(s)\, ds \right| + \left| \int_a^t \phi(s) g(s)\, ds \right| \tag{15}$$

$$\leqq \left[\phi(t)(t-a)^{1/2} + \|\phi\| \right] \|g\|.$$

Because of (6), $\phi(t)(t-a)^{1/2}$ tends to zero as $t \downarrow a$, so there exists a positive constant C such that, for any $g \in L^2(a,b)$,

$$\left| \int_a^t (\phi(t) - \phi(s)) g(s)\, ds \right| \leqq C \|g\|, \qquad t \in (a,b). \tag{16}$$

Every $y \in \operatorname{dom} M$ has a representation of the form (12), where the integral obeys the inequality (16). Clearly, y is bounded on (a,b) if and only if $c=0$.

(i) $\Leftrightarrow$ (iii). A more careful estimate of the second term in (15) yields the inequality

$$\left|\int_a^t (\phi(t)-\phi(s))g(s)\,ds\right| \leqq \left[\phi(t)(t-a)^{1/2}+\left(\int_a^t \phi^2(s)\,ds\right)^{1/2}\right]\|g\|. \tag{17}$$

Because of (6), there exists a positive constant C such that $\phi(s) \leqq C(s-a)^{-\gamma}$ for s sufficiently close to a. Thus we find that, for any $g \in L^2(a,b)$, we have the more refined estimate

$$\left|\int_a^t (\phi(t)-\phi(s))g(s)\,ds\right| \leqq \psi(t)\|g\|, \qquad t \in (a,b), \tag{18}$$

where ψ is independent of g, ψ is bounded on (a,b) and $\psi(t)=O((t-a)^{1/2-\gamma})$ as $t \downarrow a$. Every $y \in \operatorname{dom} M$ has a representation of the form (12), where the integral tends to zero like $(t-a)^{1/2-\gamma}$ as $t \downarrow a$. Clearly, $y(t)$ tends to a finite limit as $t \downarrow a$ if and only if $c=0$.

(i) $\Leftrightarrow$ (iv). The proof is similar to the proof of the previous equivalence.

(i) $\Leftrightarrow$ (v). For any $g \in L^2(a,b)$ we have

$$\left|\int_a^t g(s)\,ds\right| \leqq (t-a)^{1/2}\|g\|, \qquad t \in (a,b). \tag{19}$$

Representing y as in (12), so py' is given by (13), we see that $(py')(t)$ tends to zero as $t \downarrow a$ if and only if $c=0$.

(i) $\Leftrightarrow$ (vi). The equivalence follows immediately from the proof of the previous equivalence.

(i) $\Leftrightarrow$ (vii). According to Lemma 1, we have for any $y \in \operatorname{dom} M$,

$$p^{1/2}(t)y'(t) = -cp^{-1/2}(t) - p^{-1/2}(t)\int_a^t g(s)\,ds, \qquad t \in (a,b), \tag{20}$$

for some constant c and some $g \in L^2(a,b)$. Now, using (19),

$$\int_a^b p^{-1}(t)\left|\int_a^t g(s)\,ds\right|^2 dt \leqq \|g\|^2 \int_a^b p^{-1}(t)(t-a)\,dt.$$

The last integral is bounded:

$$\int_a^b p^{-1}(t)(t-a)\,dt = -\int_a^b \phi'(t)(t-a)\,dt = -(b-a)+\int_a^b \phi(t)\,dt = C,$$

so

$$\int_a^b p^{-1}(t)\left|\int_a^t g(s)\,ds\right|^2 dt \leqq C\|g\|^2.$$

The second term in the right member of (20) defines therefore an element of $L^2(a,b)$. The first term, on the other hand, does not, unless $c=0$. Consequently, $p^{1/2}y' \in L^2(a,b)$ if and only if $c=0$.

(i) $\Leftrightarrow$ (viii). We use the representation (20) of $p^{1/2}y'$ and observe that

$$\int_a^b (t-a)^{-\alpha}p^{-1}(t)\left|\int_a^t g(s)\,ds\right|^2 dt \leqq \|g\|^2 \int_a^b p^{-1}(t)(t-a)^{1-\alpha}dt.$$

The last integral is still bounded if $\alpha \in (0, \frac{1}{2})$, so

$$\int_a^b (t-a)^{-\alpha} p^{-1}(t) \left| \int_a^t g(s)\, ds \right|^2 dt \leqq C \| g \|^2.$$

The expression $(t-a)^{-\alpha/2} p^{-1/2}(t) \int_a^t g(s)\, ds$ defines therefore an element of $L^2(a,b)$ as long as $\alpha \in (0, \frac{1}{2})$. On the other hand, the expression $(t-a)^{-\alpha/2} p^{-1/2}(t)$ clearly does not define an element of $L^2(a,b)$ if $\alpha \in (0, \frac{1}{2})$, so the function $t \mapsto (t-a)^{-\alpha/2} p^{1/2}(t) y'(t)$ with $\alpha \in (0, \frac{1}{2})$ belongs to $L^2(a,b)$ if and only if $c = 0$.

(i) $\Leftrightarrow$ (ix). According to Lemma 1 we have, for any $y \in \operatorname{dom} M$,

$$y'(t) = -cp^{-1}(t) - p^{-1}(t) \int_a^t g(s)\, ds, \qquad t \in (a,b), \tag{21}$$

for some constant c and some $g \in L^2(a,b)$. Now,

$$\int_a^b \left| p^{-1}(t) \int_a^t g(s)\, ds \right| dt \leqq \int_a^b p^{-1}(t) \int_a^t |g(s)|\, ds\, dt$$

$$= \int_a^b |g(s)| \int_s^b p^{-1}(t)\, dt\, ds \leqq \| g \|^2 \left(\int_a^b \left(\int_s^b p^{-1}(t)\, dt \right)^2 ds \right)^{1/2}.$$

The last integral is bounded, so the second term in the right member of (21) defines an element of $L^1(a,b)$. The first term does not, unless $c = 0$. Hence, $y' \in L^1(a,b)$ if and only if $c = 0$. □

Lemma 2 shows that the domain of the maximal operator M can be restricted in many equivalent ways. Let T be defined by

$$Ty = My, \qquad y \in \operatorname{dom} T, \tag{22}$$

where $\operatorname{dom} T = \{ y \in \operatorname{dom} M : y$ satisfies (8) and any one of the conditions (i)–(ix) of Lemma 2$\}$.

THEOREM 3. *T is selfadjoint in $L^2(a,b)$.*

Proof. Let $f, g \in \operatorname{dom} T$. Then

$$(Tf, g) = [f, g]_a^b + (f, Tg),$$

where

$$[f, g] = -(pf')\bar{g} + f(p\bar{g}').$$

The bilinear form $[\cdot, \cdot]$ vanishes at b, because f and g satisfy the boundary condition (9). It also vanishes at a, as one verifies most easily using the conditions (iii) and (v) of Lemma 2. Hence, T is symmetric.

It follows from Lemma 1 and the definition of T that M, the maximal operator, is a one-dimensional extension of T. Furthermore, M is a two-dimensional extension of the minimal operator associated with τ. Since T is symmetric, we have $T^* \supset T$. But T^* cannot be a proper extension of T, because then T^* would coincide with M; hence, $T^* = T$. □

The operator T coincides with the Friedrichs extension of the minimal operator associated with the differential expression τ. Hence, we have established several equivalent characterizations of the Friedrichs extension. In the special case of the Legendre differential operator, a proof of the equivalence of the condition (iii), (v) and (vii) of Lemma 2 can be found in Akhiezer and Glazman [2, Appendix II, §9]; the other characterizations appear to be new. The simple characterization given by the condition (i) of Lemma 2 appears to be particularly interesting.

REFERENCES

[1] M. A. NAIMARK, *Linear Differential Operators*, 2 Vols., Frederick Ungar Publ., New York, 1967, 1968.

[2] N. I. AKHIEZER AND I. M. GLAZMAN, *Theory of Linear Operators in Hilbert Space*, 2 Vols., Frederick Ungar Publ. Co., New York, 1961, 1963.

[3] H. WEYL, *Über gewöhnliche Differentialgleichungen mit Singularitäten und die zugehörigen Entwicklungen willkürlicher Funktionen*, Math. Ann., 68 (1910), pp. 220–269.

[4] E. C. TITCHMARSH, *Eigenfunction Expansions Associated with Second-Order Differential Equations*, 2 Vols., 2nd Edition, Oxford Univ. Press, Cambridge, 1962.

SIAM J. MATH. ANAL.
Vol. 17, No. 4, July 1986

003

STÄCKEL-EQUIVALENT INTEGRABLE HAMILTONIAN SYSTEMS*

C. P. BOYER†, E. G. KALNINS‡ AND W. MILLER, JR.§

Abstract. The Stäckel transform is a mapping of the commuting constants of the motion (corresponding to a separable coordinate system) for one completely integrable classical or quantum Hamiltonian system to the constants of the motion for another such system. Here the transform is defined and given an intrinsic characterization, and a large family of nontrivial examples is worked out of systems which are "Stäckel equivalent". Among the simplest examples are geodesic flow on an n-dimensional ellipsoid with distinct axes, which is equivalent to the motion of a mass point on the unit sphere in R^{n+1} under the influence of a quadratic potential with distinct eigenvalues, and the Kepler (Coulomb) problem in three dimensions which is equivalent to the pseudo-Coulomb problem.

Key words. Stäckel transform, Hamiltonian system, separation of variables

AMS(MOS) subject classifications. Primary 35A22, 35J05, 58F05, 70H20

Introduction. It has long been known that the Hamiltonian systems corresponding to geodesic flow on an n-dimensional ellipsoid in R^{n+1} with distinct axes and to the motion of a mass point on the unit sphere in R^{n+1} under the influence of a quadratic potential with distinct eigenvalues are both completely integrable [1], [2]. The original proofs were based on Jacobi's separation of variable techniques for solution of the Hamilton–Jacobi equation. With the recent discovery of many examples of completely integrable Hamiltonian systems which are not separable and the attempt to develop a general theory for all such systems, interest in the old (separable) examples has revived, [3], [4]. Now, however, the emphasis is on the explicit construction of the algebraic constants of the motion for these systems in the cotangent bundle of R^{n+1}. (This was not done by Jacobi.)

Uhlenbeck [5] and Devaney [6] appear to have been the first to discover and utilize the algebraic constants of the motion for the sphere problem. In [3] Moser developed a geometric approach to the ellipsoid and sphere systems in which he was able to derive the algebraic constants of the motion for both systems without using separation of variables. He made the interesting observation that the two systems are closely related: under the "hodograph transformation" $p_i \to -x^i$, $x^i \to p_i$ the constants of the motion for one go to the constants of the motion for the other. (The relationship persists for the quantum mechanical analogies of these systems where the hodograph transformation is replaced by the Fourier transform; see the "sphere model" in [7].)

This striking correspondence between two physically distinct separable systems motivated the authors of the present work to study the relationship of these systems from the viewpoint of separation of variables. (This makes good sense since there is a one-to-one relationship between separable coordinates for a system and certain involutive families of constants of the motion for the system. Separable coordinates have a coordinate-free characterization.) Our study has led us to the concept of the Stäckel

*Received by the editors October 19, 1983, and in revised form April 10, 1985.

†Department of Mathematics and Computer Science, Clarkson College of Technology, Potsdam, New York 13676.

‡Mathematics Department, University of Waikato, Hamilton, New Zealand.

§School of Mathematics, University of Minnesota, Minneapolis, Minnesota 55455. The work of this author was partially supported by the National Science Foundation under grant MCS 82-19847.

transform, which maps the constants of the motion for one orthogonal separable system to the constants of the motion for another orthogonal separable system. Two systems related by a sequence of Stäckel transforms are said to be Stäckel equivalent. Interestingly this equivalence, while intrinsic has no general connection with the hodograph transform.

The basic idea behind the Stäckel transform is related to the Jacobi metric in analytical dynamics [8, p. 172]. Let $H(\mathbf{y},\mathbf{p})$ be a Hamiltonian, quadratic in the momenta p_i, and $V(\mathbf{y})$ a scalar potential such that the Hamilton–Jacobi equation $H+V=E$ is separable in the orthogonal coordinates $\{x^i\}$. Then the Hamilton–Jacobi equation $V^{-1}H=E$ is also separable in these coordinates for the new Hamiltonian $V^{-1}H$. This invertible transform $H\to V^{-1}H$ can be defined intrinsically, i.e., independent of coordinates, by extending it to the involutive family of constants of the motion that characterize $\{x^i\}$.

In §1 we review the intrinsic characterizations of orthogonal coordinate separation for the Hamilton–Jacobi and Helmholtz (Schrödinger) equations in terms of second order symmetries of these equations. In §2 we define the Stäckel transform between two Hamiltonian systems, first in a coordinate dependent manner and then intrinsically (Theorem 3). In §3 we work out a family of (nontrivial) examples of Stäckel equivalent systems, one of the simplest of which is the pair of systems discussed by Moser. In each case one of the systems consists of the n-dimensional geodesic flat space or constant curvature Hamiltonian with an added separable potential. The other system is either of the same type or is an induced Hamiltonian on an n-dimensional coordinate hypersurface in $(n+1)$-dimensional flat space. The explicit algebraic integrals for each of these systems are not worked out here but can be obtained directly from the results of [9].

Finally, in §4 we show that the classical mechanics ideas of §2 can be carried over to quantum mechanics without difficulty. We conclude with an intrinsic characterization of Stäckel equivalence for quantum mechanical systems.

All functions treated in this paper are assumed to be locally analytic.

1. Intrinsic characterization of separation. Let V^n be an n-dimensional (local) pseudo-Riemannian manifold and $\hat{V}^n$ the associated cotangent bundle ($n\geqq 2$). If $(y^1,\cdots,y^n)$ is a local coordinate system on V^n then $(y^1,\cdots,y^n,\ p_1,\cdots,p_n)$ is the corresponding canonical system of coordinates on $\hat{V}^n$. If the metric ds^2 on V^n takes the form

$$ds^2=\sum g_{ij}dy^i dy^j$$

on V^n, then the Hamiltonian H on $\hat{V}^n$ can be written as

$$H(\mathbf{y},\mathbf{p})=\sum g^{ij}(\mathbf{y})\,p_i p_j \tag{1.1}$$

and the Laplace–Beltrami operator is expressed as

$$\Delta=\frac{1}{\sqrt{g}}\sum \partial_i\left(\sqrt{g}\,g^{ij}\partial_j\right) \tag{1.2}$$

where $g=\det(g_{ij})$, and $\sum_j g^{ij}g_{jk}=\delta^i_k$.

The *Hamilton–Jacobi equation* is

$$H\left(y^i,\partial_j S\right)\equiv\sum_{i,j} g^{ij}\partial_i S\partial_j S=E \tag{1.3}$$

and the *Helmholtz equation* is

$$\Delta\psi(y^i)=E\psi(y^i). \tag{1.4}$$

Our interest is in additively separable orthogonal coordinate systems for (1.3) and multiplicatively R-separable orthogonal coordinate systems for (1.4). We know that the orthogonal coordinate system $\{x^i\}$ is additively separable for (1.3) if and only if the metric

$$ds^2=\sum_i H_i^2(dx^i)^2 \tag{1.5}$$

is in Stäckel form [10] (see (2.1)). Furthermore, there is the following intrinsic characterization of variable separation [11], [12]:

THEOREM 1. *Necessary and sufficient conditions for the existence of an orthogonal separable coordinate system* $\{x^i\}$ *for the Hamilton–Jacobi equation* (1.3) *are that there exist n quadratic functions* $G_k=\sum_{i,j} g^{ij}_{(k)} p_i p_j$, $(G_1=H)$ *on* $\hat{V}^n$ *such that*

1) $\{G_k, G_l\}=0$, $1\leqq k, l\leqq n$.

2) *The set* $\{G_k\}$ *is linearly independent* (*as n quadratic forms*).

3) *There is a basis* $\{\omega_{(j)}: 1\leqq j\leqq n\}$ *of simultaneous eigenforms for the* $\{G_k\}$.

(*Here we follow the definitions of Eisenhart* [13] *for the roots and eigenforms of a quadratic form with respect to a metric.*)

If conditions 1)–3) are satisfied, then there exist functions $g^j(x)$ such that $\omega_{(j)}=g^j dx^j$, $j=1,\cdots,n$. (Here $\{\ ,\ \}$ is the Poisson bracket on V^n.) The separable solutions of (1.3) are characterized by the equations $G_k(x^i,\partial_j S)=\lambda_j$, $j=1,\cdots,n_j$ $(\lambda_1=E)$ where the λ_j are the separation parameters.

Recall that a function $V(\mathbf{x})$ is a *Stäckel multiplier* with respect to a Stäckel form metric $ds^2=\sum H_i^2(dx^i)^2$ provided the metric $d\hat{s}^2=V ds^2$ is also in Stäckel form [14].

PROPOSITION. *Let* $ds^2=\sum H_i^2(dx^i)^2$ *be a Stäckel form metric. The following are equivalent*:

1) $V(\mathbf{x})$ *is a Stäckel multiplier.*

2) *There exist functions* $\psi_j=\psi_j(x^j)$ *such that*

$$V(\mathbf{x})=\sum_{j=1}^{n}\psi_j(x^j)H_j^{-2}.$$

3) *The function* $V(\mathbf{x})$ *satisfies*

$$\partial_{jk}V-\partial_j V\partial_k\log H_j^{-2}-\partial_k V\partial_j\log H_k^{-2}=0, \qquad j\neq k, \qquad j,k=1,\cdots,n. \tag{1.6}$$

Here $\partial_j=\partial_{x^j}$.

In this paper we shall be concerned with the Hamilton–Jacobi equation with potential,

$$\sum_{i,j} g^{ij}(\mathbf{y})\partial_i S\partial_j S+V(\mathbf{y})=E. \tag{1.7}$$

The technical conditions for additive separation of this equation in the orthogonal coordinates $\{x^i\}$ are that a) the metric ds^2, (1.5), is in Stäckel form with respect to these coordinates and b) the potential V is a Stäckel multiplier in the coordinates $\{x^i\}$.

Theorem 1 extends to the following intrinsic characterization of variable separation for (1.7) [15].

THEOREM 2. *Necessary and sufficient conditions for the existence of an orthogonal additive separable coordinate system* $\{x^i\}$ *for the Hamilton–Jacobi equation with potential* $H(\mathbf{y}, \partial_i S)+V(\mathbf{y})=E$ *are that there exist* n *quadratic functions* $G_k=\sum_{i,j} h^{ij}_{(k)} p_i p_j$, $(G_1=H)$ *and* n *functions* $V_k(\mathbf{y})$, $(V_1=V)$ *on* $\hat{V}^n$ *such that*

1) $\{G_k+V_k, G_l+V_l\}=0$, $1\leqq k, l\leqq n$.

2) *The set* $\{G_k\}$ *is linearly independent* (*as n quadratic forms*).

3) *There is a basis* $\{\omega_{(j)}: 1\leqq j\leqq n\}$ *of simultaneous eigenforms for the* $\{G_k\}$.

The separable solutions are characterized by the equations $G_k(x^i_j \partial_j S)+V_k(\mathbf{x})=\lambda_k$, $k=1, \cdots, n$, $(\lambda_1=E)$ *where the* λ_k *are the separation parameters.*

Here the eigenforms $\omega_{(j)}$ are related to the separable coordinates dx^j by $\omega_{(j)}=g^j(\mathbf{x})dx^j$ where g^j is a nonzero function. The above results are easily proved by slight modifications of the proof of [16, Thm. 3].

2. Stäckel equivalence. Let H be a quadratic Hamiltonian on a $2n$-dimensional symplectic manifold, $\{x^i\}$ be an orthogonal separable coordinate system for the Hamilton–Jacobi equation $H(x^i, \partial_j S)=E$ and $V(\mathbf{x})$, $W(\mathbf{x})$ be two nonzero Stäckel multipliers for H in the coordinates $\{x^i\}$. Then the Hamilton–Jacobi equation

$$H\left(x^i, \partial_j S\right)+V(\mathbf{x})=E$$

separates in the coordinates $\{x^i\}$ and so does the equation

$$H'\left(x^i, \partial_j S\right)+W^{-1}V=E$$

where $H'=W(x)^{-1}H$ is a Hamiltonian on a new manifold $\hat{V}^n$. (Indeed, $W^{-1}V$ is a Stäckel multiplier for H'.) We say that $H'+W^{-1}V$ is a *Stäckel transform* of $H+V$.

We first examine the significance of Stäckel transforms in terms of Stäckel matrices. Now $H=\sum_{i=1}^n H_i^{-2} p_i^2$ is in Stäckel form, i.e., there exists an $n\times n$ invertible matrix $S=(S_{ji}(x^j))$ such that

$$H_j^{-2}=(S^{-1})^{nj}, \qquad j=1, \cdots, n \tag{2.1}$$

where S^{-1} is the multiplicative inverse of S. Since W is a Stäckel multiplier for H there exist functions $l_i(x^i)$ such that

$$W(\mathbf{x})=\sum_{j=1}^{n} l_j(x^j) H_j^{-2}. \tag{2.2}$$

It follows easily that $W^{-1}H$ is in Stäckel form with $n\times n$ Stäckel matrix

$$S'=\begin{pmatrix} & l_1(x^1) \\ S_{j\alpha}(x^j) & \vdots \\ & l_n(x^n) \end{pmatrix}, \qquad 1\leqq j\leqq n, \quad 1\leqq \alpha\leqq n-1. \tag{2.3}$$

Thus the Stäckel transform consists in the replacement of the nth column of S by the column vector $(l_1(x^1), \cdots, l_n(x^n))$. (We have singled out the nth column by convention. More generally, we shall see that the replacement of any column in a Stäckel matrix by a new column vector $(l_1(x^1), \cdots, l_n(x^n))$ defines a Stäckel transform so long as the new matrix is nonsingular, i.e., so long as the corresponding potential is nonzero.)

A more sophisticated approach to the Stäckel transform is through the roots $\rho_k^{(i)}$ of the quadratic functions G_i which characterize the separable coordinates, Theorems 1

and 2. As is well known, the symmetries G_i take the form

$$G_i = \sum_{j=1}^{n} \rho_j^{(i)} H_j^{-2} p_j^2, \qquad i = 1, \cdots, n \tag{2.4}$$

where $G_1 = H$, $\rho_j^{(1)} = 1$ and, ([16], [13, Appendix 13])

$$\partial_q \rho_k^{(i)} = \left(\rho_q^{(i)} - \rho_k^{(i)}\right) \partial_q \log H_k^{-2}, \qquad q, k = 1, \cdots, n. \tag{2.5}$$

If $W(\mathbf{x}) = \sum_{j=1}^{n} l_j(x^j) H_j^{-2}$ is a Stäckel multiplier for H, then it is straightforward to show that the functions

$$W_i(\mathbf{x}) = \sum_{j=1}^{n} l_j(x^j) \rho_j^{(i)} H_j^{-2} \tag{2.6}$$

satisfy the conditions $\{G_i + W_i, G_k + W_k\} = 0$, $W_1 = W$ and characterize the separable coordinates $\{x^i\}$, as shown in Theorem 2.

A set of roots $\rho_k^{(i)'}$ corresponding to the transformed Hamiltonian $W^{-1}H$ is given by

$$\rho_k^{(i)'} = W \rho_k^{(i)} - W_i + 1, \qquad i = 1, \cdots, n. \tag{2.7}$$

This follows easily from

LEMMA 1. *Let $(\rho_1(x), \cdots, \rho_n(x))$ be a solution of the system*

$$\partial_q \rho_k = (\rho_q - \rho_k) \partial_q \log H_k^{-2}, \qquad q, k = 1, \cdots, n \tag{2.8}$$

where $H = \sum_{i=1}^{n} H_i^{-2} p_i^2$ is in Stäckel form. Then

$$\rho_k' = \sum_{i=1}^{n} l_i(x^i)(\rho_k - \rho_i) H_i^{-2} + 1, \qquad k = 1, \cdots, n$$

is a solution of the system

$$\partial_q \rho_k' = \left(\rho_q' - \rho_k'\right) \partial_q \log\left(H_k^{-2}/W\right) \tag{2.9}$$

where $W = \sum_{i=1}^{n} l_i(x^i) H_i^{-2}$.

Finally, the Stäckel transform has an intrinsic characterization.

THEOREM 3. *Let $(G_i, V_i : i = 1, \cdots, n)$ be a set of quadratic polynomials and potentials corresponding to an orthogonal coordinate system $\{x^i\}$ for the Hamiltonian $H + V \equiv G_1 + V_1$ $(V \not\equiv 0)$, i.e.,*

1) $\{G_k + V_k, G_l + V_l\} = 0$, $1 \leqq k, l \leqq n$.

2) *The set of quadratic forms $\{G_k\}$ is linearly independent.*

3) *There is a basis $\{\omega_{(j)} : 1 \leqq j \leqq n\}$ of simultaneous eigenforms for the $\{G_k\}$.*

Suppose the system (G_i, W_i) also satisfies conditions 1)–3) *with the same basis of eigenforms. Then the system (G_i', V_i') where*

$$\begin{aligned} G_j' &= G_j - W_j W_1^{-1} G_1 + W_1^{-1} G_1, \\ V_j' &= V_j - W_j V_1 / W_1 + V_1 / W_1, \qquad j = 1, \cdots, n \end{aligned} \tag{2.10}$$

satisfies conditions 1)–3) *and corresponds to the orthogonal coordinate system $\{x^i\}$ for the Hamiltonian $G' + V' \equiv G_1' + V_1'$. In particular, $\{G_k' + V_k', G_l' + V_l'\} = 0$.*

Proof. It is clear that the system (G_i', V_i'), defined by (2.10), satisfies conditions 2) and 3). Condition 1) for the systems (G_i, V_i), (G_i, W_i) is equivalent to the requirements

a) $\{G_k, G_l\} = 0$,

b) $\{G_k, V_l\} + \{V_k, G_l\} = 0$, $\{G_k, W_l\} + \{W_k, G_l\} = 0$, $1 \leqq k,\ l \leqq n$.

It is straightforward to show that a) and b) together with the properties $\{L, M_1 + M_2\} = \{L, M_1\} + \{L, M_2\}$, $\{L, M\} = -\{M, L\}$ and $\{L, M_1 M_2\} = M_1\{L, M_2\} + \{L, M_1\} M_2$ of the Poisson bracket for functions $L(\mathbf{x}, \mathbf{p})$, $M_i(\mathbf{x}, \mathbf{p})$ imply $\{G_1', G_j'\} = 0$ and $\{G_j', G_l'\} = 0$, $2 \leqq j,\ l \leqq n$. Furthermore $\{G_k', V_l'\} + \{V_k', G_l'\} = 0$. Q.E.D.

The Stäckel transform is invertible in the following sense.

COROLLARY 1. *Let H, G_i, V_i, V and G_i', V_i', W, W_i be defined as in Theorem* 3. *Let $U = W^{-1}$. Then U is a Stäckel multiplier for $G_1' + V_1' = W^{-1}H + W^{-1}V$ in the orthogonal separable coordinates $\{x^i\}$, corresponding to $(G_i',\ V_i' : i = 1, \cdots, n)$. Furthermore the transformed system is identical to (G_i, V_i).*

For computational convenience the preceding results have been expressed in terms of an explicit basis G_i of quadratic forms and an associated basis V_i, W_i of potentials. However, the results are clearly basis free: they apply to an n-dimensional vector space $\mathscr{K}$ of quadratic forms K in involution, and associated potentials $V_K(\mathbf{x})$, $W_K(\mathbf{x})$. The maps $K \to V_K$, $K \to W_K$ are linear. Furthermore we can extend the definition of the Stäckel transform by noting that any nonsingular quadratic form $\mathscr{K}_0 \in \mathscr{K}$ for which $W_{K_0} \not\equiv 0$ can serve as a *Hamiltonian* with which to define a transform. Thus a given space $\mathscr{K}$ is associated with many different Hamilton–Jacobi equations. It is this simple observation which leads to the most interesting applications of the Stäckel transform.

Let $\mathscr{K}$ and $\mathscr{L}$ be two n-dimensional vector spaces of quadratic forms in involution, each of which has a simultaneous eigenbasis of differential forms. We say that $\mathscr{L}$ is *Stäckel equivalent* to $\mathscr{K}$ if there is a finite sequence of Stäckel transforms $a_1, \cdots, a_t$ such that $\mathscr{L} = a_t \circ a_{t-1} \circ \cdots \circ a_1(\mathscr{K})$. (If b is a Stäckel transform of $\mathscr{K}$ and a is a Stäckel transform of $b(\mathscr{K})$ then $a \circ b$ *is the composition which maps* $\mathscr{K}$ *to* $a(b(\mathscr{K}))$.) We are associating the potential $V_K \equiv 0$ with each of $\mathscr{K}$ and $\mathscr{L}$. It is clear from Corollary 1 that Stäckel equivalence is a true equivalence relation.

THEOREM 4. *Let $\mathscr{K}$ and $\mathscr{L}$ be n-dimensional vector spaces of quadratic forms in involution, each of which has a simultaneous eigenbasis of differential forms. Then $\mathscr{L}$ is Stäckel equivalent to $\mathscr{K}$ if and only if $\mathscr{K}$ and $\mathscr{L}$ have the same eigenbasis.*

Proof. Suppose $\mathscr{L}$ and $\mathscr{K}$ are Stäckel equivalent. Since a space of quadratic forms and its Stäckel transform have the same eigenforms, it follows that $\mathscr{L}$ and $\mathscr{K}$ must have the same eigenforms.

Conversely, suppose $\mathscr{K}$ and $\mathscr{L}$ have the same eigenforms. It follows from Theorem 1 that there exist coordinates $(x^1, \cdots, x^n)$ such that the basis of eigenforms is $(dx^1, \cdots, dx^n)$. Thus $\mathscr{K}$ and $\mathscr{L}$ correspond to $n \times n$ Stäckel matrices $S_{\mathscr{K}}$ and $S_{\mathscr{L}}$ in the same coordinates. Let $\mathbf{k}_i(x^i)$, $\mathbf{l}_i(x^i)$ be the ith column vector in $S_{\mathscr{K}}$, $S_{\mathscr{L}}$, respectively. Note that an elementary column transformation of $S_{\mathscr{K}}$ in which column $\mathbf{k}_i$ is replaced by $\mathbf{k}_i + \alpha \mathbf{k}_h$, where α is a constant and $h < i$, leads to another Stäckel matrix associated to the same Hamiltonian $K = K_1$ and the same coordinates. (Indeed such a transformation merely corresponds to a change of basis forms K_j, $j = 2, \cdots, n$.) The same remarks hold for $S_{\mathscr{L}}$.

By using the elementary column transformations if necessary, we can always choose $S_{\mathscr{K}}$ and $S_{\mathscr{L}}$ such that the following properties hold: each matrix S_j, $0 \leqq j \leqq n$ where $S_j = (\mathbf{l}_1(x^1), \mathbf{l}_2(x^2), \cdots, \mathbf{l}_j(x^j), \mathbf{k}_{j+1}(x^{j+1}), \cdots, \mathbf{k}_n(x^n))$, $j = 1, \cdots, n-1$, $S_0 = S_{\mathscr{L}}$, $S_n = S_{\mathscr{K}}$ is nonsingular and each matrix element $(S^{-1})_j^{ih}$, $(0 \leqq j \leqq n,\ 1 \leqq i, h \leqq n)$ is nonzero. From the remarks following (2.3) we see that the replacement $S_j \to S_{j-1}$ defines a Stäckel transform a_j and $\mathscr{L} = a_1 \circ \cdots \circ a_n(\mathscr{K})$. Q.E.D.

COROLLARY 2. *If $\mathscr{L}$ and $\mathscr{K}$ are Stäckel equivalent then $\mathscr{K}$ can be mapped to $\mathscr{L}$ by a sequence of at most n Stäckel transforms.*

It is a consequence of Theorem 4 that (roughly speaking) *any* orthogonal separable coordinate system on an n-dimensional pseudo-Riemannian manifold is Stäckel equivalent to any such system on another n-manifold. The primary practical and theoretical interest of Stäckel equivalence concerns systems that are equivalent via a single Stäckel transform.

The following construction, a special case of the general Stäckel transform, constitutes an important application of transform methods to the study of Hamiltonian systems. Let H be a quadratic Hamiltonian on a $2n$-dimensional symplectic manifold $\hat{V}_n$ and suppose the Hamilton–Jacobi equation

$$H(x^i, \partial_j S) - \lambda W(\mathbf{x}) + V(\mathbf{x}) = E, \qquad W \not\equiv 0 \tag{2.11}$$

separates in the orthogonal coordinates $\{x^i\}$ for all values of the parameter λ. Then $H(x^i, p_j) = \sum_{i=1}^n H_i^{-2} p_i^2$ is in Stäckel form and the potentials W, V are Stäckel multipliers for H. Furthermore, via Theorem 3, this separable system is characterized by n linearly independent quadratic forms $G_k(\mathbf{x}, \mathbf{p})$ on $\hat{V}_n$ and $2n$ potentials $W_k(\mathbf{x}), V_k(\mathbf{x})$, $k = 1, \cdots, n$ such that $G_1 \equiv H$, $W_1 \equiv W$, $V_1 \equiv V$ and the family $\{G_k - \lambda W_k + V_k\}$ is in involution with respect to the Poisson bracket. The separable complete integral $S(\mathbf{x}; \lambda; \mathbf{E}) = \sum S_i(x^i; \lambda; \mathbf{E})$ of (2.11) is characterized by the equations

$$G_k(x^i; \partial_j S) - \lambda W_k(x) + V_k(x) = E_k, \qquad k = 1, \cdots, n \tag{2.12}$$

with $E_1 = E$. Dividing (2.11) by the nonzero Stäckel multiplier $W(\mathbf{x})$, we obtain the transformed Hamilton–Jacobi equation

$$W^{-1}H - EW^{-1} + W^{-1}V = \lambda \tag{2.13}$$

where now E is considered as a parameter and λ is the "energy". Clearly (2.13) admits the same separable complete integral $S(\mathbf{x}; \lambda; \mathbf{E})$ as does (2.11), with separation parameters $\lambda, E_2, \cdots, E_n$. Furthermore the separable solutions are now characterized by the Hamiltonian $W_1^{-1}G_1 - EW_1^{-1} + W_1^{-1}V_1$ and the constants of the motion

$$(G_k - W_1^{-1}W_k G_1) + EW_1^{-1}W_k + (V_k - V_1 W_1^{-1} W_k), \qquad k = 2, \cdots, n. \tag{2.14}$$

This corresponds to a special case of Theorem 5.

This special Stäckel transform can be very useful when both the systems (2.12) and (2.13), (2.14) correspond to spaces of physical or geometrical interest: one of the systems may prove more tractable than the other. In particular, the Stäckel transform does not, in general, preserve the symmetry algebra of a Hamilton–Jacobi equation so one system may have a higher degree of symmetry than the other.

In the following section we will derive a number of examples of Hamiltonian systems on flat spaces and spaces of constant curvature that are related by Stäckel transforms.

3. Examples. We begin the section with some simple low-dimensional cases where Stäckel transforms appear and are useful. This will be followed by a derivation of families of examples associated with constant curvature spaces in n dimensions.

One of the simplest examples is associated with the Coulomb problem [17]. The Hamilton–Jacobi equation for this problem (expressed in Cartesian coordinates x^i) is

$$p_1^2 + p_2^2 + p_3^2 - \frac{q}{r} = E, \qquad r = \left[\sum_{i=1}^{3} (x^i)^2 \right]^{1/2}. \tag{3.1}$$

Here q is a real parameter. This equation separates in spherical coordinates (r,θ,φ). The associated pseudo-Coulomb problem

$$r\left(p_1^2+p_2^2+p_3^2\right)-Er=q \tag{3.2}$$

also separates in these coordinates, where now E is considered as a parameter and q is the "energy". Here (3.2) is obtained as a Stäckel transform of (3.1) by the potential r^{-1}. Although (3.1) and (3.2) are equivalent, (3.2) is more tractible from a symmetry point of view. Indeed for (3.2) the fundamental constants of the motion (the angular momentum vector and the Laplace–Runge–Lenz vector) form a 6-dimensional symmetry algebra isomorphic to $O(4)$. However for (3.1) the corresponding constants of the motion fail to generate a finite-dimensional algebra under the Poisson bracket [17]. Similar comments apply for the quantum Kepler problem [18].

For our next example we use the fact that if $ds^2=\sum_{i=1}^n H_i^2(dx^i)^2$ is in Stäckel form and $\partial_{x^n}H_j=0$ for all j then H_n^{-2} is a Stäckel multiplier for the reduced Hamiltonian associated with the $(n-1)$-dimensional space $ds'^2=\sum_{i=1}^{n-1} H_i^2(dx^i)^2$ and coordinates $x^1,\cdots,x^{n-1}$. (This follows directly from the Levi–Civita conditions [13, p. 265].) Let $H=p_1^2+p_2^2+p_3^2$ be the Hamiltonian for three-dimensional Euclidean space (expressed in Cartesian coordinates x^i) and pass to parabolic coordinates

$$x^1=\xi\eta\cos\theta,\ x^2=\xi\eta\sin\theta,\qquad x^3=\tfrac{1}{2}\left(\xi^2-\eta^2\right). \tag{3.3}$$

The Hamilton–Jacobi equation for these orthogonal separable coordinates takes the form

$$H=\frac{1}{\xi^2+\eta^2}\left(p_\xi^2+p_\eta^2\right)+\frac{1}{\xi^2\eta^2}p_\theta^2=E. \tag{3.4}$$

Since θ is an ignorable variable, corresponding to a complete separable integral we have $p_\theta=\lambda$, a separation parameter. Now the potential $\xi^{-2}\eta^{-2}$ is a Stäckel multiplier for the reduced Hamiltonian $(\xi^2+\eta^2)^{-1}(p_\eta^2+p_\xi^2)$. Dividing by this multiplier, we can recast $H=E_1$ in the form

$$\frac{\xi^2\eta^2}{\xi^2+\eta^2}\left(p_\xi^2+p_\eta^2\right)+\xi^2\eta^2p_\varphi^2=-\lambda^2=E' \tag{3.5}$$

where $p_\varphi=\sqrt{-E}$. (Here, we must make the restriction that both E and E' are negative.) Now (3.5) is just the Hamilton–Jacobi equation for the hyperboloid

$$t^2-x^2-y^2-z^2=1,\qquad t>1. \tag{3.6}$$

Indeed the coordinates are related by

$$x=\frac{\varphi}{\xi\eta},\quad y=\frac{1}{2}\left(\frac{\xi}{\eta}-\frac{\eta}{\xi}\right),\quad t+z=\frac{1}{\xi\eta},\quad t-z=\frac{\left(\xi^2+\eta^2\right)^2}{4\xi\eta}+\frac{\varphi^2}{\xi\eta}. \tag{3.7}$$

Note that the Hamilton–Jacobi equations for these two geometrically distinct problems have essentially the same solutions.

For a more sophisticated example we consider the Hamiltonian $H=p_1^2+p_2^2-\lambda(z_1^2+z_2^2)$ on Euclidean two-space, expressed in terms of the Cartesian coordinates z_1, z_2. In terms of elliptic coordinates x_1, x_2 the Hamilton–Jacobi equation is separable

and the separation is characterized by

(3.8)

$$H=G_1=-4\left(\frac{(e_1-x_1)(e_2-x_1)}{x_1-x_2}p_{x_1}^2+\frac{(e_1-x_2)(e_2-x_2)}{x_2-x_1}p_{x_2}^2\right)+\lambda(x_1+x_2)=E_1,$$

$$G_2=-4\left(\frac{x_2(e_1-x_1)(e_2-x_1)}{x_1-x_2}p_{x_1}^2+\frac{x_1(e_1-x_2)(e_2-x_2)}{x_2-x_1}p_{x_2}^2\right)+\lambda x_1x_2=E_2.$$

Here $e_1<e_2$ are real constants and for fixed z_1, z_2 with $z_1z_2\neq 0$ the corresponding elliptic coordinates x_i are uniquely chosen such that $x_1<e_1<x_2<e_2$ and $x_i=\mu$ is a solution of

$$\frac{z_1^2}{e_1-\mu}+\frac{z_2^2}{e_2-\mu}=1.$$

Now consider the case $e_1=e$, $e_2=0$ and note that x_1x_2 is a Stäckel multiplier for G_2. Dividing by this multiplier, we can write $G_2=E_2$ in the form

$$H'=4\left(\frac{x_1-e}{x_1-x_2}p_{x_1}^2+\frac{x_2-e}{x_2-x_1}\right)p_{x_2}^2+\frac{E_2}{x_1x_2}=\lambda. \tag{3.9}$$

This is just the Hamilton–Jacobi equation, expressed in parabolic coordinates, corresponding to two-dimensional Euclidean space with an added potential. In Cartesian coordinates z_i, $H'=\lambda$ where

$$H'=p_1^2+p_2^2+\frac{E_2}{2ez_2-z_1^2}. \tag{3.10}$$

Here for given z_1, z_2 with $z_1\neq 0$ the corresponding parabolic coordinates x_i are uniquely chosen such that $x_1<e<x_2<0$ and $x_i=\mu$ is a solution of

$$2z_2-\mu+\frac{z_1^2}{\mu-e}=0.$$

For the case $e_1e_2\neq 0$ it is still true that x_1x_2 is a Stäckel multiplier for G_2. Dividing by the multiplier we can rewrite $G_2=E_2$ in the form

$$H''=-4\left(\frac{(e_1-x_1)(e_2-x_1)}{x_1(x_1-x_2)}p_{x_1}^2+\frac{(e_1-x_2)(e_2-x_2)}{x_2(x_2-x_1)}p_{x_2}^2\right)-\frac{E_2}{x_1x_2}=-\lambda. \tag{3.11}$$

This is the Hamilton–Jacobi equation in parabolic coordinates induced on the paraboloid

$$2z_3-\frac{z_1^2}{e_1}-\frac{z_2^2}{e_2}=0$$

in three-dimensional Euclidean space by the potential $E_2(z_1^2+z_2^2-2z_3(e_1+e_2)-e_1e_2)^{-1}$.

Since many of our examples concern coordinate hypersurfaces corresponding to orthogonal separable coordinate systems, we will briefly digress to discuss the induced Riemannian structure on such hypersurfaces. Let V^n be a local pseudo-Riemannian manifold and $\hat{V}^n$ the associated cotangent bundle. Let H be the Hamiltonian on $\hat{V}^n$

and suppose the Hamilton–Jacobi equation $H(x^i, \partial_j S) = E$ separates in the orthogonal coordinates $(x^1, \cdots, x^n)$. By Theorem 1 these coordinates can be associated to an involutive family $\{G_i : i = 1, \cdots, n\}$ of quadratic forms on $\hat{V}^n$ where $H = G_1$. The metric, expressed in the separable coordinates, is

$$ds^2 = \sum_{j=1}^{n} H_j^2(\mathbf{x})(dx^j)^2. \tag{3.12}$$

Now consider the pseudo-Riemannian manifold Z^{n-1} obtained by restricting V^n to a coordinate hypersurface, say $x^n = c$, c a constant. The induced metric on Z^{n-1} is

$$d\tilde{s}^2 = \sum_{j=1}^{n=1} H_j^2(\tilde{\mathbf{x}}, c)(dx^j)^2, \qquad \tilde{\mathbf{x}} = (x^1, \cdots, x^{n-1}). \tag{3.13}$$

Similarly, we can identify the cotangent bundle $\hat{Z}^{n-1}$ as a restriction of $\hat{V}^n$. If $(\mathbf{x}, \mathbf{p})$ are local symplectic coordinates on $\hat{V}^n$, then $(\tilde{\mathbf{x}}, \tilde{\mathbf{p}})$ are local coordinates on $\hat{Z}^{n-1}$, where $\tilde{\mathbf{p}} = (p_1, \cdots, p_{n-1})$ and for the restriction we set $x^n = c$, $p_n = 0$. The pullback of the quadratic forms G_i to $\hat{Z}^{n-1}$ is

$$\tilde{G}_i = \sum_{j=1}^{n-1} \rho_j^{(i)}(\tilde{\mathbf{x}}, c) H_j^{-2}(\tilde{x}, c) p_j^2, \qquad i = 1, \cdots, n \tag{3.14}$$

and $\{\tilde{G}_i, \tilde{G}_l'\} = 0$, $1 \leqq i$, $l \leqq n$, where $\{\cdot, \cdot\}'$ is the Poisson bracket on $\hat{Z}^{n-1}$. The Hamilton–Jacobi equation for Z^{n-1}

$$\tilde{H}(\tilde{x}, \partial_j W) = H(\tilde{\mathbf{x}}, c, \partial_j W, 0) = E, \qquad j = 1, \cdots, n-1, \tag{3.15}$$

separates in the coordinates $(x^1, \cdots, x^{n-1})$ and these coordinates are uniquely associated with the $(n-1)$-dimensional involutive family spanned by the forms $\tilde{G}_i$, $i = 1, \cdots, n$.

We see that an orthogonal separable coordinate system $\{x^1, \cdots, x^n\}$ for V^n leads to orthogonal separable systems and corresponding involutive families of quadratic forms for all of the pseudo-Riemannian manifolds induced on the coordinate hypersurfaces.

Our basic example is the "generic" separable system $\{x^1, \cdots, x^n\}$ on V^n with metric

$$ds^2 = \sum_{j=1}^{n} \frac{\prod_{i \neq j}(x^i - x^j)(dx^j)^2}{f(x^j)} \tag{3.16}$$

where f is a nonzero function of a single variable. (At this point we shall not be precise about the range of values permitted to the x^j.) It is evident that this system is separable. Indeed the columns of the Stäckel matrix can be chosen as

$$\mathbf{l}_i(\mathbf{x}) = \begin{pmatrix} \dfrac{(x^1)^{i-1}}{f(x^1)} \\ \vdots \\ \dfrac{(x^n)^{i-1}}{f(x^n)} \end{pmatrix}, \qquad i = 1, \cdots, n. \tag{3.17}$$

It is not difficult to show [9] that V^n is a space of nonzero constant curvature if and only if $(d^{n+1}/du^{n+1})f(u)=c$, a nonzero constant. Furthermore, V^n is flat if and only if $(d^{n+1}/du^{n+1})f(u)=0$. If V^n is Euclidean, then f is a polynomial of order $n-1$ or n.

Note that the Hamiltonian in these coordinates can be taken to be

$$G_1=\sum_{j=1}^{n}\frac{f(x^j)}{\Pi_{i\neq j}(x^i-x^j)}p_j^2 \tag{3.18}$$

and the quadratic form obtained by expanding the Stäckel matrix on the 1st column is

$$G_n=\sum_{j=1}^{n}\frac{f(x^j)\Pi_{i\neq j}(x^i)}{\Pi_{i\neq j}(x^i-x^j)}p_j^2. \tag{3.19}$$

For convenience we choose the coordinate hypersurface Z^{n-1} in the normalized form: $x^n=0$. Then the metric on Z^{n-1} in the separable coordinates $(x^1,\cdots,x^{n-1})$ is

$$d\tilde{s}^2=-\sum_{j=1}^{n-1}\frac{x^j\Pi_{i\neq j}(x^i-x^j)(dx^j)^2}{f(x^j)} \tag{3.20}$$

and the Hamiltonian is

$$\tilde{G}_1=-\sum_{j=1}^{n-1}\frac{f(x^j)}{x^j\Pi_{i\neq j}(x^i-x^j)}p_j^2 \tag{3.21}$$

where now $i<n$. Note that $\tilde{G}_n\equiv 0$.

Making use of the identities

$$\sum_{l=1}^{n}\frac{(x^l)^k}{\Pi_{i\neq l}(x^l-x^i)}=\begin{cases}0, & k=0,1,\cdots,n-2,\\ 1, & k=n-1,\\ \sum_{l=1}^{n}x^l, & k=n,\end{cases} \tag{3.22}$$

for the case $k=n$, we see that the potential

$$V_1=-\sum_{j=1}^{n}x^j \tag{3.23}$$

is a Stäckel multiplier for the Hamiltonian G_1. Furthermore, using (3.11) for $k=n-1$, we see that the potential V_n associated with V_1 is

$$V_n=-x^1x^2\cdots x^n. \tag{3.24}$$

It follows that V_n is a Stäckel multiplier for G_n and the Stäckel transform

$$K=V_n^{-1}G_n=-\sum_{j=1}^{n}\frac{f(x^j)}{x^j\Pi_{i\neq j}(x^i-x^j)}p_j^2 \tag{3.25}$$

is the Hamiltonian for the coordinate hypersurface Z^n in V_{n+1}, with metric (3.16), same f, n replaced by $n+1$ (compare with (3.21)). We have shown that (for all f) the system generated by the Hamiltonian G_1, (3.18), in V^n is Stäckel equivalent to the system generated by the Hamiltonian K, (3.25), in Z^n. We can consider Z^n as an n-dimensional hypersurface in V^{n+1}.

All of the examples where f is a polynomial of order $\leqq n+1$ occur as hypersurfaces in $(n+1)$-dimensional flat space. Well known is the case [9], [10]

$$f(u)=(u-e_1)(u-e_2)\cdots(u-e_{n+1}), \qquad e_1<e_2<\cdots<e_{n+1}. \tag{3.26}$$

Here the embedding space is Euclidean space E^{n+1} with Cartesian coordinates $\mathbf{z}=(z_1,\cdots,z_{n+1})$ and metric

$$ds^2=dz_1^2+\cdots+dz_{n+1}^2. \tag{3.27}$$

We now consider the unit sphere S^n in E^{n+1},

$$z_1^2+z_2^2+\cdots+z_{n+1}^2=1. \tag{3.28}$$

The metric on S^n is obtained by restricting ds^2 to (3.28). The appropriate coordinates on S^n are elliptic spherical coordinates $(x^1,\cdots,x^n)$ defined as follows, [3], [9]: through each point $\hat{\mathbf{z}}\in S^n$ with no $\hat{z}_i=0$ there pass exactly n confocal cones

$$\sum_{i=1}^{n+1}\frac{z_i^2}{e_i-\lambda}=0, \tag{3.29}$$

corresponding to the parameters $\lambda=x^1,\cdots,x^n$ where

$$e_1<x^1<e_2<x^2<e_3<\cdots<x^n<e_{n+1}. \tag{3.30}$$

Indeed we have the identity in λ

$$\sum_{i=1}^{n+1}\frac{z_i^2}{e_i-\lambda}=\frac{\prod_{j=1}^{n}(x^j-\lambda)}{\prod_{l=1}^{n+1}(e_l-\lambda)}. \tag{3.31}$$

The metric on S^n expressed in the elliptic spherical coordinates x^i takes the form $-\frac{1}{4}$ times (3.16) with f given by (3.26) [9]. Comparing coefficients of λ^{-2} in the Laurent series expansions for both sides of (3.31), we obtain the result

$$V_1=-\sum_{j=1}^{n}x^j=\sum_{i=1}^{n+1}e_iz_i^2-\sum_{l=1}^{n+1}e_l. \tag{3.32}$$

Since the constant Σe_l can be absorbed into E for the Hamilton–Jacobi equation $G_1+V_1=E$, we can consider the separable potential V_1 as a quadratic potential in Cartesian coordinates.

Now we consider elliptic coordinates in E^{n+1}. Through each point $\mathbf{z}\in E^{n+1}$ with no $z_i=0$ there pass exactly $n+1$ confocal ellipsoids [3], [19]

$$\sum_{i=1}^{n+1}\frac{z_i^2}{e_i-\lambda}=1 \tag{3.33}$$

corresponding to the parameters $\lambda=x^0,\cdots,x^n$ where

$$x^0<e_1<x^1<e_2<\cdots<x^n<e_{n+1}. \tag{3.34}$$

We have the identity in λ

$$1-\sum_{i=1}^{n+1}\frac{z_i^2}{e_i-\lambda}=\prod_{j=1}^{n+1}\frac{(\lambda-x^{j-1})}{(\lambda-e_j)}. \tag{3.35}$$

Suppose $e_1 > 0$ and consider the ellipsoid Z^n obtained by setting $x^0 = 0$. It is straightforward via (3.14) to show that the induced metric on this hypersurface is

$$d\tilde{s}^2 = -\frac{1}{4}\sum_{j=1}^{n} \frac{x^j \Pi_{i\neq j}(x^i - x^j)(dx^j)^2}{f(x^j)}$$

where f is given by (3.26). This shows that the Hamiltonian system on the sphere S^n (elliptic spherical coordinates) is Stäckel equivalent to the Hamiltonian system on the ellipsoid (elliptic coordinates). (Moser [3] has demonstrated that these two systems are equivalent via the hodograph transformation, see also [4]. Similarly, the "sphere model" in [7, Chap. 3] displays the equivalence of these systems for the Helmholtz equation. However, the following Stäckel equivalent systems are not equivalent via the hodograph transform.)

We now consider the cases where $\deg f = n+1$ and f has repeated roots, or $\deg f \leqq n$. In each case we will compute the expression of the separable potential $V_1 = -\Sigma_{j=1}^n x^j$ in terms of "natural" cartesian coordinates $\{z^i\}$ and identify the coordinate hypersurfaces in terms of cartesian coordinates. The method we follow, a limit procedure based on (3.31), was pioneered by Bôcher [19] and refined by two of the authors [20]. The results are complete except that determination of the range of the coordinates is left to the reader. (The "natural" coordinates z_i, y_i may be complex but the separable coordinates and the roots will be real.)

Consider first the case $\dim f = n+1$, where f has roots e_J of multiplicities $N_1, \cdots, N_P$, $\Sigma_{J=1}^P N_J = n+1$. Then, as shown in [17], the identity (3.31) is modified to

$$\sum_{J=1}^{P}\sum_{j=1}^{N_J} \frac{S_{j+1}^J}{(\lambda - e_J)^{N_J+1-j}} = \frac{\Pi_{j=1}^n(\lambda - x^j)}{\Pi_{J=1}^P(\lambda - e_J)^{N_J}} \tag{3.36}$$

where $S_{j+1}^J = \Sigma_i Y_i^J Y_{j+1-i}^J$. Here

$$\sum_i z_i^2 = \sum_J S_{N_J+1}^J = \sum_J \sum_j \left(z_j^J\right)^2 = 1$$

where

(i) $N_J = 2p$ (even),

$$Y_j^J = \frac{1}{\sqrt{2}}\left(z_j^J - iz_{N_J+1-j}^J\right), \qquad Y_{2P+1-j}^J = \frac{1}{\sqrt{2}}\left(z_j^J + iz_{N_J+1-j}^J\right), \qquad j = 1, \cdots, p,$$

(ii) $N_J = 2p+1$ (odd), as above but $Y_{p+1}^J = z_{p+1}^J$.

Here the coordinates Y_i^J are real. Equating coefficients of λ^{-2} on both sides of (3.36), we obtain

$$\sum_j x^j - \sum_J N_J e_J = \sum_{J=1}^{P}\left(S_{N_J}^J + e_J S_{N_J+1}^J\right). \tag{3.37}$$

It follows from (3.26) that the coordinate curves are given by

$$\sum_{J=1}^{P}\sum_{j=1}^{N_J} \frac{S_{j+1}^J}{(\lambda - e_j)^{N_J+1-j}} = 0. \tag{3.38}$$

We can take $z_1 = z_1^1$, $z_2 = z_2^1, \cdots, z_{N_1} = z_{N_1}^1$, $z_{N_1+1} = z_1^2, \cdots, z_n = z_{N_P}^P$.

For example, if $N_1=2$, and $N_j=1$ for $1<j\leqq n-1$ we have

$$\sum_j x^j-2e_1-\sum_{j=2}^{n} e_j=e_1\left(z_1^2+z_2^2\right)+\frac{1}{2}\left(z_1+iz_2\right)^2+\sum_{j=2}^{n} e_j z_{j+1}^2.$$

To treat the case $\deg f\leqq n$, we start from the basic identity [19], [20]

$$\sum_{j=1}^{n+2}\frac{y_j^2}{(\lambda-e_j)}=\left(-\sum_{j=1}^{n+2} e_j y_j^2\right)\frac{\prod_{j=1}^{n}(\lambda-x^j)}{\prod_{l=1}^{n+2}(\lambda-e_l)} \tag{3.39}$$

for general cyclidic coordinates x^j. Here the y_j are hyperspherical coordinates ($\Omega\equiv\sum_{j=1}^{n+2}y_j^2=0$), related to cartesian coordinates z_i by formulas

$$z_i=y_i/\left(\sqrt{-1}\,y_{n+1}+y_{n+2}\right),\qquad i=1,\cdots,n. \tag{3.40}$$

(Note: There is some freedom in the choice of y_{n+1}, y_{n+2}.) Now suppose that $n+2$ roots e_l degenerate to roots with multiplicities $N_1,\cdots,N_P$ where $\sum_J N_J=n+2$. As shown in [20], formula (3.39) then becomes

$$\sum_{J=1}^{P}\sum_{j=1}^{N_J}\frac{S_{j+1}^J}{(\lambda-e_J)^{N_J+1-j}}=\left(-\sum_J e_J S_{N_J+1}^J\right)\frac{\prod_{j=1}^{n}(\lambda-x^j)}{\prod_{J=1}^{P}(\lambda-e_J)^{N_J}} \tag{3.41}$$

and $\Omega=\sum_J S_{N_J+1}^J=0$. At this point it is convenient to set $\lambda=-1/\lambda'$, $x^j=-1/x'^j$, $e_J=-1/e_J'$ so that (3.41) becomes

$$\sum_{J=1}^{P}\sum_{j=1}^{N_J-1}\frac{S_{j+1}^J}{(\lambda'-e_J')^{N_J+1-j}}(\lambda')^{N_J-1-j}\left(e_J'\right)^{N_J+1-j}+\sum_{J=1}^{P}\frac{S_{N_J+1}^J}{(\lambda'-e_J')} \tag{3.42}$$

$$=\frac{\left(\sum_J\left(1/e_J'\right)S_{N_J+1}^J\right)}{\left(\prod_{j=1}^{n}x'^j\right)}\cdot\frac{\left(\prod_{J=1}^{P}\left(e_J'\right)^{N_J}\right)}{\prod_{j=1}^{P}(\lambda'-e_J)^{N_J}}\prod_{j=1}^{n}(\lambda'-x'^j).$$

Here we have added $\lambda'\Omega$ to the left-hand side of this equation and restricted λ' to the domain $|\lambda'|>|e_j|$, $\forall j$.

Following Bôcher's procedures, to obtain the coordinate curves corresponding to f with $\deg f\leqq n$ and distinct roots of multiplicities $N_2,\cdots,N_P$ respectively, we let $e_1'\to\infty$ in (3.42). (Here, we are assuming $N_1\geqq 2$.) We use the notation $[\overset{\infty}{N}_1,N_2,\cdots,N_P]$ to denote this coordinate system. The result is easily seen to be

(3.43)

$$\sum_{j=1}^{N_1-1}S_{j+1}^1(-\lambda')^{N_1-j-1}+\sum_{J=2}^{P}\sum_{j=1}^{N_J-1}\frac{S_{j+1}^J}{(\lambda'-e_J')^{N_J-j+1}}\left(\lambda' e_J'\right)^{N_J-j+1}\lambda'^{-2}+\sum_{J=2}^{P}\frac{S_{N_J+1}^J}{(\lambda'-e_J')}$$

$$=\frac{\left(\sum_{J=2}^{P}\left(1/e_J'\right)S_{N_J+1}^J\right)}{\prod_{j=1}^{n}x'^j}\frac{\left(\prod_{J=2}^{P}\left(e_J'\right)^{N_J}\right)}{\prod_{J=2}^{P}\left(\lambda'-e_J'\right)^{N_J}}(-1)^{N_1}\prod_{j=1}^{n}(\lambda'-x'^j).$$

Equating coefficients of $(\lambda')^{N_1-2}$ on each side of this identity, we obtain

$$S_2^1=\frac{\left(\sum_{J=2}^{P}(1/e_J')S_{N_J+1}^J\right)\left(\prod_{J=2}^{P}(e_J')^{N_J}\right)}{\prod_{j=1}^{n}x'^j} \tag{3.44}$$

Equating coefficients of $(\lambda')^{N_1-3}$ and making use of the condition $\Omega=0$, we have

$$\sum_{j=1}^{n}x'^j-\sum_{J=2}^{P}N_Je_J'=\begin{cases}\dfrac{S_3^1}{S_2^1}, & N_1\geqq 3,\\[2ex] \dfrac{S_3^1}{S_2^1}, & N_1=2.\end{cases} \tag{3.45}$$

Here $S_2^1=(y_1^1)^2$, $S_3^1=2y_1^1y_2^1$.

For $N_1>3$ we can define hyperspherical coordinates $\tilde{y}_j^J$ so that

$$\sqrt{2}\,y_1^1=\tilde{y}_1^1-\sqrt{-1}\,\tilde{y}_{N_1}^1,\qquad \sqrt{2}\,y_2^1=\tilde{y}_2^1-\sqrt{-1}\,\tilde{y}_{N_1-1}^1 \tag{3.46}$$

and (of course)

$$\sum_{J=1}^{P}\sum_{j=1}^{N_J}\left(\tilde{y}_j^J\right)^2=0.$$

Passing to Cartesian coordinates z_j^J via (3.40), we can obtain the linear potential

$$\frac{S_3^1}{S_2^1}=\sqrt{2}\left(z_2^1-\sqrt{-1}\,z_{N_1-1}^1\right).$$

This potential will be real only for a pseudo-Euclidean space.

For $N_1=3$ we can set $y_2^1=\tilde{y}_2^1$ and, from (3.46), (3.40), achieve the linear potential

$$\frac{S_3^1}{S_2^1}=\sqrt{2}\,z_2^1.$$

Finally, for $N_1=2$ we can obtain the oscillator potential

$$\frac{S_3^1}{S_2^1}=\sum_{J=2}^{P}\sum_{j=1}^{N_J}\left(z_j^J\right)^2=\sum_{l=1}^{n}(z_l)^2.$$

Thus all the potentials V_1 correspond to either linear or oscillator potentials in terms of Cartesian coordinates. The coordinate hypersurface $x=x^i$ for $[\overset{\infty}{N}_1,N_2,\cdots,N_P]$ is given by

$$\sum_{j=1}^{N_1-1}S_{j+1}^1(-x^i)^{N_1-j-1}+\sum_{J=2}^{P}\sum_{j=1}^{N_J-1}\frac{S_{j+1}^J}{\left(x^i-e_J\right)^{N_J-j+1}}\left(x^ie_J\right)^{N_J-j+1}(x^i)^{-2}$$
$$+\sum_{J=2}^{P}\frac{S_{N_J+1}^J}{\left(x^i-e_j\right)}=0. \tag{3.47}$$

Note that since $S^1_{N_1+1}$ does not occur explicitly in (3.47), hyperspherical coordinates y_l can always be chosen such that this hypersurface is quadratic in terms of the associated cartesian coordinates in E^{n+1}.

4. Stäckel equivalence for Schrödinger equations. Here we work out the appropriate modification of the results of §2 for the Schrödinger (or Helmholtz) equation with velocity dependent potential. On V^n this equation can be written in the form

$$\mathscr{H}\psi \equiv \Delta\psi + Q\cdot\nabla\psi + V\psi = E\psi \tag{4.1}$$

where Δ is the Laplacian

$$\Delta = \frac{1}{\sqrt{g}} \sum_{i,j=1}^{n} \partial_i\left(\sqrt{g}\, g^{ij}\partial_j\right),$$

and

$$Q\cdot\nabla\psi = \sum_{i,j=1}^{n} Q_i g^{ij}\partial_j\psi. \tag{4.2}$$

Here, $Q_i(\mathbf{y})$ is a covariant vector and $V(\mathbf{y})$ is a scalar function. Applying the definitions and techniques of [16], [21], we obtain the following necessary and sufficient conditions for multiplicative R-separability of (4.1) in the orthogonal coordinates $\{x^i\}$, $(\psi = e^R\prod_{j=1}^n \psi^{(j)}(x^j))$:

1) the metric $ds^2 = \sum_i H_i^2 (dx^i)^2$ is in Stäckel form,

2) there is a function Q such that $Q_i = \partial_i Q$ in the coordinates $\{x^j\}$,

3) $\sum_{i=1}^n H_i^{-2}(R_{ii} - 2R_i^2) + V$ is a Stäckel multiplier where

$$-2R_i = \partial_i \log\left(H_1\cdots H_n/H_i^2\right) + Q_i, \qquad R_{ii} = \partial_i R_i. \tag{4.3}$$

If conditions 1)–3) are satisfied, then $R_i = \partial_i R$.

Now suppose that the Schrödinger equation (4.1) R-separates in the orthogonal coordinates $\{x^j\}$. In these coordinates (4.1) takes the form

$$\sum_i \left\{ \frac{1}{\sqrt{g}}\partial_i\left[\left(\sqrt{g}\, H_i^{-2}\right)\partial_i\right] + Q_i H_i^{-2}\partial_i \right\}\psi + V\psi = E\psi, \tag{4.4}$$

where $\sqrt{g} = H_1 H_2 \cdots H_n$. Let $W \neq 0$ be a Stäckel multiplier for the metric $ds^2 = \sum H_i^2 (dx^i)^2$. Our basic observation is that the Schrödinger equation

$$W^{-1}[\Delta + Q\cdot\nabla + V]\psi = E\psi \tag{4.5}$$

i.e.,

$$[\Delta' + Q'\cdot\nabla + W^{-1}V]\psi = E\psi \tag{4.6}$$

is also R-separable in the coordinates $\{x^j\}$. Here Δ' is the Laplace operator corresponding to the metric $W\,ds^2$. We say that $\Delta' + Q'\cdot\nabla + W^{-1}V$ is a *Stäckel transform* of $\Delta + Q\cdot\nabla + V$.

Variable R-separation for (4.1) can be characterized intrinsically.

THEOREM 5. *Necessary and sufficient conditions for the existence of an orthogonal R-separable coordinate system $\{x^i\}$ for the Schrödinger equation* (4.1) *are that there exist n second-order differential operators*

$$\mathscr{A}_k = \Delta_k + \sum_{i,j=1}^{n} \partial_{yi} Q g^{ij}_{(k)} \partial_{yj} + V_k, \qquad k = 1,\cdots,n, \tag{4.7}$$

($\mathscr{A}_1=\mathscr{H}$) on V^n where

$$\Delta_k=\frac{1}{\sqrt{g}}\sum_{i,j=1}^{n}\partial_{yi}\left(\sqrt{g}\,g^{ij}_{(k)}\partial_{yj}\right) \tag{4.8}$$

and $\Delta_1=\Delta$ such that

1) $[\mathscr{A}_k,\mathscr{A}_l]\equiv\mathscr{A}_k\mathscr{A}_l-\mathscr{A}_l\mathscr{A}_k=0$, $1\leqq k,l\leqq n$.

2) *The set $\{G_k: k=1,\cdots,n\}$ is linearly independent as n quadratic forms on $\hat{V}^n$, where*

$$G_k=\sum_{i,j=1}^{n}g^{ij}_{(k)}p_ip_j.$$

3) *There is a basis $\{\omega_{(j)}: 1\leqq j\leqq n\}$ of simultaneous eigenforms for the $\{G_k\}$.*

If conditions 1)–3) *are satisfied, then there exist functions $g^j(\mathbf{x})$ such that $\omega_{(j)}=g^j\,dx^j$.* (*Note that the division* (4.7) *of $\mathscr{H}_k$ into three terms is coordinate independent. Indeed the operators Δ_k are in self-adjoint form with respect to the measure on V^n given in local coordinates by $\sqrt{g}\,dy^1\cdots dy^n$ whereas Q and V_k $V(_1=V)$ are scalar valued functions on V^n.*)

Proof. Suppose the operators satisfy conditions 1)–3). Comparing coefficients of third derivative terms in $[\mathscr{A}_k,\mathscr{A}_l]=0$, we obtain $\{G_k,G_l\}=0$. It follows from Theorem 1 and comparison of second derivative terms that there exists an orthogonal coordinate system $\{x^i\}$ such that

$$\mathscr{A}_k=\sum_{i=1}^{n}\rho_i^{(k)}H_i^{-2}(\partial_{ii}+f_i\partial_i)+\sum_{i=1}^{n}Q_i\rho_i^{(k)}H_i^{-2}\partial_i+V_k. \tag{4.9}$$

Here $f_i=\partial_i\log(H_1\cdots H_n/H_i^2)$, $\partial_i=\partial_{x^i}$, $Q_i=\partial_{x^i}Q$ and the functions $\rho_i^{(k)}(\mathbf{x})$ satisfy conditions (2.4), (2.5).

At this point it is convenient to perform the similarity transformation

$$\mathscr{A}_k\leftrightarrow\tilde{\mathscr{A}}_k=e^{-R}\mathscr{A}_ke^{R} \tag{4.10}$$

where $-2\partial_iR=f_i+Q_i$. We find

$$\tilde{\mathscr{A}}_k=\sum_i\rho_i^{(k)}H_i^{-2}\partial_{ii}+\tilde{V}_k \tag{4.11}$$

where

$$\tilde{V}_k=V_k+\sum_i\rho_i^{(k)}H_i^{-2}\left(R_{ii}-2R_i^2\right). \tag{4.12}$$

Comparison of the coefficients of ∂_l in $[\tilde{\mathscr{A}}_1,\tilde{\mathscr{A}}_k]=0$ leads to

$$\partial_l\tilde{V}_k=\partial_l\left(\rho_l^{(k)}T\right),\qquad 2\leqq k\leqq n \tag{4.13}$$

where

$$T=V+\sum_iH_i^{-2}\left(R_{ii}-2R_i^2\right). \tag{4.14}$$

Since $\partial_{lj}\tilde{V}_k = \partial_{jl}\tilde{V}_k$ we see from (2.5) and (4.13) that

$$\partial_{jl}T = \partial_j T \partial_l \log H_j^{-2} + \partial_l T \partial_j \log H_l^{-2}, \qquad j \neq l.$$

Thus T is a Stäckel multiplier and (4.1) R-separates in the orthogonal coordinates $\{x^i\}$.

Conversely, if (4.1) R-separates in the orthogonal coordinates $\{x^i\}$ then there exist functions $\psi_l(x^l)$ such that $T = \sum_l \psi_l(x^l) H_l^{-2}$ where T is given by (4.14). It is straightforward to verify that the operators $\mathscr{A}_k = e^R \tilde{\mathscr{A}}_k e^{-R}$ defined by (4.11), (4.12) with

$$\tilde{V}_k = \sum_l \rho_l^{(k)} \psi_l(x^l) H_l^{-2},$$

satisfy conditions 1)–3). Q.E.D.

Shapovalov [15] has stated without proof a theorem which implies Theorem 5.

COROLLARY 3. *Let* $\{\mathscr{A}_k : k = 1, \cdots, n\}$ *satisfy conditions* 1)–3) *of Theorem* 5 *and let* $\{W_k : k = 1, \cdots, n\}$ *be a set of scalar valued functions on* V^n *such that* $W_1 \neq 0$ *and*

$$\{G_k + W_k, G_l + W_l\} = 0, \qquad 1 \leqq k, l \leqq n. \tag{4.15}$$

Then the system $\{\mathscr{A}'_k\}$ *where*

$$\mathscr{A}'_j = \mathscr{A}_j - W_j W_1^{-1} \mathscr{A}_1 + W_1^{-1} \mathscr{A}_1, \qquad j = 1, \cdots, n$$

also satisfies conditions 1)–3) *and corresponds to the orthogonal coordinate system* $\{x^i\}$ *for the Schrödinger operator* $\mathscr{H}' = \mathscr{A}'_1$. *In particular* $\{\mathscr{A}'_k, \mathscr{A}'_l\} = 0$.

This is the operator analogue of Theorem 3 and is proved in a similar manner.

In analogy to (2.12) there is a special case of the Stäckel transform for Schrödinger equations in which the formal eigenfunctions are preserved by the transform. Consider the eigenvalue equations $\mathscr{A}_k \psi = E_k \psi$, $k = 1, \cdots, n$, corresponding to the R-separable equation ($E_1 = E$)

$$\frac{1}{\sqrt{g}} \sum_{i=1}^{n} \partial_i \left(g^{ii} \sqrt{g}\, \partial_i \right) \psi + (V - \lambda W)\psi = E\psi, \tag{4.16}$$

in the orthogonal coordinates $\{x^j\}$. Here λ is a parameter and

$$\mathscr{A}_k = \frac{1}{\sqrt{g}} \sum_{i=1}^{n} \partial_i \left(a_{(k)}^{ii} \sqrt{g}\, \partial_i \right) + V_k - \lambda W_k. \tag{4.17}$$

Suppose the quadratic form $\{a_{(l)}^{ii}\}$ is nonsingular and $W_l \neq 0$. Then W_l is a Stäckel multiplier for this form and the equation

$$\frac{1}{W_k \sqrt{g}} \sum \partial_i \left(a_{(l)}^{ii} \sqrt{g}\, \partial_i \right) \psi + \left(\frac{V_l}{W_l} - \frac{E}{W_l} \right) \psi = \lambda \psi \tag{4.18}$$

admits the same R-separable solutions as does (4.16). Here (4.18) can be cast in the self-adjoint from

$$\frac{1}{\sqrt{a W_l^n}} \sum_{i=1}^{n} \partial_i \left(\frac{a_{(l)}^{ii}}{W_l} \sqrt{a W_l^n}\, \partial_i \right) \Theta + (\tilde{V} - E\tilde{W})\Theta = \lambda \Theta \tag{4.19}$$

where $a^{-1}=\prod_{j=1}^{n} a_{(l)}^{jj}$, $\tilde{W}=W_l^{-1}$,

$$\tilde{V}=\sum_{j=1}^{n}\left(\frac{a_{(l)}^{jj}}{W_l}\left[S_{jj}+S_j^2+\partial_j \ln\left(\sqrt{g}\cdot S_j\right)\right]+\frac{\partial_j a_{(l)}^{jj}}{W_l}S_j\right) \tag{4.20}$$

and

$$\psi=e^S\Theta, \qquad e^S=\left[\frac{W_l^{n-2}a}{g}\right]^{1/4}. \tag{4.21}$$

Thus the Stäckel transform takes R-separable solutions ψ of (4.16) on the manifold V^n to R-separable solutions Θ of (4.19) on a manifold $V^{n'}$.

Note that the natural metric on V^n with respect to which the $\mathscr{A}_k$ are formally self-adjoint is $\sqrt{g}\,d\mathbf{x}$. Similarly the natural metric on $V^{n'}$ is $\sqrt{W_l^n a}\,d\mathbf{x}$. A formal computation yields

$$\int\psi\bar{\psi}\sqrt{g}\,d\mathbf{x}=\int\Theta\bar{\Theta}\frac{\sqrt{W_l^n a}}{W_l}\,d\mathbf{x}$$

or $(W_l\psi,\psi)=\langle\Theta,\Theta\rangle$ where $(\cdot,\cdot)$, and $\langle\cdot,\cdot\rangle$ are the inner products on V^n, $V^{n'}$ respectively. Thus the inner product, hence the spectral resolution for the operators $\mathscr{A}_k$ is in general not preserved by a Stäckel transform. However, in particular cases the spectrum is preserved, e.g., the transform from the Coulomb to the pseudo-Coulomb problem where the Virial Theorem [18] shows that the discrete spectrum is unchanged.

REFERENCES

[1] E. JACOBI, *Vorlesungen über Dynamik*, Gesammelte Werke, Supplement band, Berlin, 1884.

[2] C. NEUMANN, *De problemate quodam mechanico, quod ad primam integralium ultraellipticorum classem revocatur*, J. Reine Angew. Math., 56 (1859), pp. 46–63.

[3] J. MOSER, *Various aspects of integrable Hamiltonian systems*, in Dynamical Systems, Birkhauser, Boston, MA, 1980.

[4] M. ADLER AND P. VAN MOERBEKE, *Completely integrable systems, Euclidean Lie algebras and curves*, Adv. Math., 38 (1980), pp. 267–317.

[5] K. UHLENBECK, *Minimal 2-spheres and tori in S^k*, informal preprint, 1975.

[6] R. DEVANEY, *Transversal homoclinic orbits in an integrable system*, Amer. J. Math., 100 (1978), pp. 631–642.

[7] W. MILLER, JR., *Symmetry and Separation of Variables*, Addison-Wesley, Reading, MA, 1977.

[8] D. LAUGWITZ, *Differential and Riemannian Geometry*, Academic Press, New York, 1965.

[9] E. G. KALNINS AND W. MILLER, *Separation of variables on n-dimensional Riemannian manifolds* 1. *The n-sphere S_n and Euclidean n-space R_n*, 2. *The n-dimensional hyperboloid H_n*, J. Math. Phys., to appear.

[10] L. P. EISENHART, *Separable systems of Stäckel*, Ann. Math., 35 (1934), pp. 284–305.

[11] E. G. KALNINS AND W. MILLER, JR., *Killing tensors and variable separation for Hamilton–Jacobi and Helmholtz equations*, this Journal, 11 (1980), pp. 1011–1026.

[12] V. N. SHAPOVALOV, *Stäckel spaces*, Siberian Math. J., 20 (1980), pp. 790–800.

[13] L. P. EISENHART, *Riemannian Geometry*, second printing, Princeton Univ. Press, Princeton, NJ, 1949.

[14] E. G. KALNINS AND W. MILLER, JR., *Killing tensors and nonorthogonal variable separation for Hamilton–Jacobi equations*, this Journal, 12 (1981), pp. 617–639.

[15] V. N. SHAPOVALOV, *Separation of variables in second-order linear differential equations*, Differential Equations, 10 (1981), pp. 1212–1220.

[16] E. G. KALNINS AND W. MILLER, JR., *The theory of orthogonal R-separation for Helmholtz equations*, Adv. Math., 51 (1984), pp. 91–106.

[17] W. THIRRING, *A Course in Mathematical Physics*. 1. *Classical Dynamical Systems*, E. Harrell, trans., Springer-Verlag, New York, 1978. (See Section 4.2)

[18] _____, *A Course in Mathematical Physics*. 3. *Quantum Mechanics of Atoms and Molecules*, E. Harrell, trans., Springer-Verlag, New York, 1981. (See Section 4.1)

[19] M. BÔCHER, *Die Reihentwickelungen der Potentialtheorie*, B. G. Teubner, Leipzig, 1894.

[20] E. G. KALNINS, W. MILLER, JR. AND G. J. REID, *Separation of variables for complex Riemannian spaces of constant curvature*. I, Proc. Roy. Soc. Lond., A 394 (1984), pp. 183–206.

[21] E. G. KALNINS AND W. MILLER, JR., *Intrinsic characterization of variable separation for the partial differential equations of mechanics*, Proc. IUTAM-ISIMM Symposium on Modern Developments in Analytical Mechanics, Turin, 1982, pp. 511–533; Atti della Academia delle Scienze di Torino, Supplemento al Vol. 117, 1983.

SIAM J. MATH. ANAL.
Vol. 17, No. 4, July 1986

004

INTEGRABILITY OF KLEIN–GORDON EQUATIONS*

PETER A. CLARKSON†, J. BRYCE McLEOD‡, PETER J. OLVER§ AND ALFRED RAMANI¶

Abstract. Using the Painlevé test, it is shown that the only integrable nonlinear Klein–Gordon equations $u_{xt}=f(u)$ with f a linear combination of exponentials are the Liouville, sine-Gordon (or sinh-Gordon) and Mikhailov equations. In particular, the double sine-Gordon equation is not integrable.

Key words. completely integrable, Painlevé property, Klein–Gordon equation

AMS(MOS) subject classifications. Primary 35Q20, 34A20

In [7], two of the present authors (J. B. M. and P. J. O.) considered the problem of which nonlinear Klein–Gordon equations

$$u_{xt}=f(u) \tag{1}$$

are completely integrable. They referred to the Ablowitz–Ramani–Segur (ARS) conjecture, [2], [3] which states that if a partial differential equation is integrable by the inverse scattering transform (IST) method, then all its reductions to ordinary differential equations have the Painlevé property, i.e., all their moveable singularities are poles. It was shown in [7] that if $f(u)$ is a linear combination of exponentials, the only equations of type (1) whose corresponding ordinary differential equation for travelling wave solutions

$$u(x,t)=w(\xi)=w(x-ct),$$

arising from the invariance of (1) under the group of translations, has the Painlevé property, are those of the form

$$u_{xt}=c_2e^{2\beta u}+c_1e^{\beta u}+c_{-1}e^{-\beta u}+c_{-2}e^{-2\beta u} \tag{2}$$

for constants $c_2,\cdots,c_{-2}$. In fact the singularities of u are not really poles, but rather "pure logarithms" in the sense that u_x, u_t and $\exp(\beta u)$ have only poles. This extension was included in the ARS conjecture as originally stated.

A paradox apparently remained; namely that the form (2), which does include the well-known Liouville equation (only one nonzero c_i), the sine-Gordon equation ($c_2=c_{-2}=0$, $c_1=-c_{-1}$, $\beta=i$) and the Mikhailov equation ($c_1=c_{-2}=0$), [8], [9], [12] all of which are known to be completely integrable, also includes the double sine-Gordon equation ($c_2=-c_{-2}$, $c_1=-c_{-1}$, $\beta=i$), which is *not* integrable. Indeed numerical studies have shown that its travelling wave solutions do not behave like solitons under collisions, [1]. This apparent problem, however, is easily resolved if one considers a second one-parameter group of symmetries of (1),

$$(x,t)\to(\lambda x,\lambda^{-1}t), \qquad \lambda>0,$$

*Received by the editors July 3, 1984, and in revised form January 21, 1985.
† Department of Mathematics, Heriot-Watt University, Edinburgh EH14 4AS, Scotland.
‡ Mathematical Institute, Oxford University, Oxford OX1 3LB, England.
§ School of Mathematics, University of Minnesota, Minneapolis, Minnesota 55455.
¶ Centre de Physique Théoretique, Ecole Polytechnique 91128 Palaiseau, France.

leading to a different form

$$u(x,t)=w(xt)=w(\xi)$$

for the group-invariant solutions. Then w satisfies the ordinary differential equation

$$\xi w''+w'=f(w) \tag{3}$$

where $\xi=xt$.

In order to apply the ARS conjecture, we need to analyze the singularities of solutions of (3) in the case f has the form (2). To eliminate logarithmic singularities, set $v=\exp(\beta w)$, so v satisfies

$$v''=\frac{v'^2}{v}-\frac{v'}{\xi}+\frac{c_2v^3+c_1v^2+c_{-1}+c_{-2}v^{-1}}{\xi}. \tag{4}$$

All second order ordinary differential equations with the Painlevé property have been classified by Painlevé and Gambier and can be reduced, through a change of variables, to one of fifty canonical forms—see [5]. An obvious candidate to reduce (4) to is the equation

$$w''=\frac{w'^2}{w}-\frac{w'}{z}+\frac{\alpha w^2+\beta}{z}+\gamma w^3+\frac{\delta}{w}, \tag{5}$$

which is canonical form number 13 in [5, p. 335]. Thus we need to determine when (4) can be reduced to the canonical form (5). The change of variables

$$\xi=z^p, \qquad v=z^qw \tag{6}$$

reduces (4) to

$$w''=\frac{w'^2}{w}-\frac{w'}{z}+\sum b_nw^{n+1}z^{nq+p-2}, \tag{7}$$

where the sum is on $n=2,1,-1$ and -2 (but not 0!) and the b_n's and c_n's are related by irrelevant powers of p.

In order for (7) to agree with (5), we need to have all of the following four conditions to hold:

a) either $b_2=0$ or $2q+p-2=0$;
b) either $b_1=0$ or $q+p-2=-1$;
c) either $b_{-1}=0$ or $-q+p-2=-1$;
d) either $b_{-2}=0$ or $-2q+p-2=0$.

Clearly this is not possible if all the b_n's are nonzero. If only one b_n is nonzero, there are no difficulties. The original equation was the Liouville equation $u_{xt}=e^{\beta u}$, which is integrable using the Bäcklund transformation

$$w_x=u_x+\exp\frac{\beta}{2}(u+w), \qquad w_t=-u_t-\frac{2}{\beta}\exp\frac{\beta}{2}(u-w)$$

with w satisfying $w_{xt}=0$, [4]. Alternatively, we can set

$$u=\frac{1}{\beta}\log\left[2v_xv_te^v/(e^v-1)^2\beta\right],$$

leading to $v_{xt}=0$. If $b_2=b_{-2}=0$, $b_1b_{-1}\neq 0$ then (4) is already in the canonical form (6) with $\gamma=\delta=0$, $\alpha\beta\neq 0$, so no change of variable is required, i.e., $p=1$, $q=0$. This is the case of the sine (and sinh-) Gordon equations, which are integrable by inverse scattering methods. If $b_1=b_{-1}=0$, $b_2b_{-2}\neq 0$, then we again have the sine-Gordon equation, but we have made a different choice for defining v in terms of u. This should not alter the Painlevé character of the equation, and indeed $p=2$, $q=0$ will satisfy conditions a)–d). Curiously enough this reduces (4) to a canonical form (5) with $\alpha=\beta=0$, $\gamma\delta\neq 0$, which is *not* the same as above. This shows that the same equation can be reduced by different changes of variables to *different* canonical forms.

If $b_1=b_{-2}=0$, $b_2b_{-1}\neq 0$ (or, respectively, $b_{-1}=b_2=0$, $b_1b_{-2}\neq 0$), then we have the Mikhailov equation

$$u_{xt}=be^{2\beta u}+b'e^{-\beta u},$$

which was shown to be integrable by a 3×3 matrix scattering problem, [8], [9], [12]. In this case, conditions a)–d) have the solution $p=\frac{4}{3}$, $q=\frac{1}{3}$ (respectively $p=\frac{4}{3}$, $q=-\frac{1}{3}$), and hence this reduction of Mikhailov's equation does have the Painlevé property.

Finally, if $b_1b_2\neq 0$, even if $b_{-2}=b_{-1}=0$, one would need $p=0$, $q=1$ for a)–d) to be satisfied. But this is not an acceptable change of variables as ξ would not really depend on z. Thus we cannot reduce (4) to the canonical form (5) if $b_1b_2\neq 0$ whatever the values of b_{-1} and b_{-2}. By symmetry, the same holds if $b_{-1}b_{-2}\neq 0$ no matter what values b_1 and b_2 have. Of course, this does not completely prove that (4) in this case does not have the Painlevé property since (6) is not the only possible choice of change of variables and it may be possible to reduce (4) to some other canonical form. Indeed, we have just seen that starting with $b_1=b_{-1}=0$, $b_2b_{-2}\neq 0$, the change of variables (6) with $p=2$, $q=0$ is a rather contrived way to reduce (4) to the canonical form (5) compared with the more obvious choice $v=w^2$. To check that if $b_1b_2\neq 0$, equation (4) does not have the Painlevé property, one could study the behavior of its singularities and show that they are not pure poles, or, alternatively, follow through Painlevé's deviation of the fifty canonical forms, as in [5], and see that it does not fall into one of these categories.

Instead of doing this, however, it is just as easy to check the Painlevé property for the partial differential equation (2) directly, using the method introduced by Weiss et al. [10], [11], and improved by Kruskal [6]. First set $v=\exp(\beta u)$, so (2) becomes

$$vv_{xt}=v_xv_t+c_2v^4+c_1v^3+c_{-1}v+c_{-2}. \tag{8}$$

Suppose $v(x,t)$ is singular along the curve

$$\psi(x,t)=x+\varphi(t)=0$$

with φ arbitrary. Let us expand v near this curve in a Laurent series

$$v=\psi^r\sum_{n=0}^{\infty}\alpha_n(t)\psi^n. \tag{9}$$

Without loss of generality, we can suppose $c_2\neq 0$. (If $c_2=0$, $c_{-2}\neq 0$ change variables by replacing v by $1/v$; if $c_2=c_{-2}=0$, change v to v^2 if $c_1\neq 0$ and v^{-2} if only $c_{-1}\neq 0$.) Balancing the lowest powers of ψ in both sides of (8), we have one possible solution $r=-1$. Equating the coefficients of ψ^{-4}, we get

$$2\alpha_0^2\frac{d\varphi}{dt}=\alpha_0^2\frac{d\varphi}{dt}+c_2\alpha_0^4,$$

so

$$c_2\alpha_0^2=\frac{d\varphi}{dt}. \tag{10}$$

Substituting (9) into (8) and identifying the coefficients of ψ^{n-4} gives an equation for all of the α_n's except α_2 which does not appear when one equates the coefficients of ψ^{-2}. Indeed $n=2$ is the "resonance" in the ARS terminology, [3]. More precisely, the coefficient of ψ^{-3} is

$$2\alpha_0\alpha_1\frac{d\varphi}{dt}-\alpha_0\frac{d\alpha_0}{dt}=-\alpha_0\frac{d\alpha_0}{dt}+4c_2\alpha_0^3\alpha_1+c_1\alpha_0^3,$$

which by (10) gives the expression

$$\alpha_1=-c_1/2c_2 \tag{11}$$

for α_1. At order ψ^{-2}, we find

$$2\alpha_0\alpha_2\frac{d\varphi}{dt}-\alpha_1\frac{d\alpha_0}{dt}=-2\alpha_0\alpha_2\frac{d\varphi}{dt}+6c_2\alpha_0^2\alpha_1^2+4c_2\alpha_0^3\alpha_2+3c_1\alpha_0^2\alpha_1.$$

By (10), (11) all the terms cancel except for $\alpha_1 d\alpha_0/dt$, the value of which is

$$\alpha_1\frac{d\alpha_0}{dt}=-\frac{c_1}{4c_2^{3/2}}\frac{d^2\varphi/dt^2}{\sqrt{d\varphi/dt}}.$$

If this quantity does not vanish, one cannot find an expansion for v in the form (9). Terms of the form

$$\psi(\alpha_2+\tilde{\alpha}_2\log\psi)$$

are needed at that order, and at higher and higher orders in ψ one will need higher and higher powers of logarithms of ψ. Such an expansion is not of Painlevé type.

In [7] the proof of the ARS conjecture was done by first showing that the solution $v(x,t)$ must be meromorphic when both x and t are allowed to assume complex values. This was then specialized to gain the required Painlevé property of group-invariant solutions. It also immediately gives the modification of the ARS conjecture by Weiss et al. [10], [11] which predicts that an equation will not be integrable if some nonpolar singularity exists on a line $\psi(x,t)=x+\varphi(t)=0$ for φ arbitrary. In particular, if $d^2\varphi/dt^2\neq 0$, then c_1 must vanish. This leads to an understanding of the result of [7]. For translation-invariant solutions, we have only straight lines $x=\lambda t+k$, so $d^2\varphi/dt^2=0$ in this case, and a nonvanishing α_1 does not pose any difficulty. Alternatively, the original ARS conjecture for the scale-invariant solutions would also lead to the same conclusion $c_1=0$.

So far we did not find any restrictions on c_{-1} and c_{-2}. However, if c_{-2} does not vanish, a similar argument shows that $c_{-1}=0$, namely we look at solutions with the

$$v=\alpha_0\psi+\alpha_1\psi^2+\cdots,$$

which, because of the coefficient v multiplying the highest derivative v_{xt} in (9) may be singular when $v=0$. Indeed, if $c_{-1}\neq 0$ one finds that it is a singular logarithmic point, with logarithms first entering at order ψ^3.

In conclusion the analysis of the singular behavior of solutions shows that the solutions are meromorphic along arbitrary curves if and only if $b_1b_2=0$ and $b_{-1}b_{-2}=0$. We conclude that the only possible integrable cases are the Liouville, sine-Gordon

(sinh-Gordon) and Mikhailov equations, in perfect agreement with the known integrable character of these equations and the nonintegrable character of the double sine-Gordon equation, as suggested by numerical studies of its solutions.

REFERENCES

[1] M. J. ABLOWITZ, M. D. KRUSKAL AND J. F. LADIK, *Solitary wave collisions*, SIAM J. Appl. Math., 36 (1979), pp. 428–437.

[2] M. J. ABLOWITZ, A. RAMANI AND H. SEGUR, *Nonlinear evolution equations and ordinary differential equations of Painlevé type*, Lett. Nuovo Cimento, 23 (1978), pp. 333–338.

[3] _____, *A connection between nonlinear evolution equations and ordinary differential equations of P-type* I, J. Math. Phys., 21 (1980), pp. 715–721.

[4] R. L. ANDERSON AND N. H. IBRAGIMOV, *Lie–Bäcklund Transformations in Applications*, SIAM Studies 1, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1979.

[5] E. L. INCE, *Ordinary Differential Equations*, Dover, New York, 1956.

[6] M. D. KRUSKAL, private communication.

[7] J. B. MCLEOD AND P. J. OLVER, *The connection between partial differential equations soluble by inverse scattering and ordinary differential equations of Painlevé type*, this Journal, 14 (1983), pp. 488–506.

[8] A. V. MIKHAILOV, *Integrability of a two-dimensional generalization of the Toda chain*, Soviet Phys. JETP Lett., 30 (1979), pp. 414–418.

[9] _____, *The reduction problem and the inverse scattering method*, Physica, 3D (1981), pp. 73–117.

[10] J. WEISS, M. TABOR AND G. CARNEVALE, *The Painlevé property for partial differential equations*, J. Math. Phys., 24 (1983), pp. 522–526.

[11] J. WEISS, *The Painlevé property for partial differential equations.* II: *Bäcklund transformation, Lax pairs and the Schwarzian derivative*, J. Math. Phys., 24 (1983), pp. 1405–1413.

[12] A. P. FORDY AND J. D. GIBBONS, *Integrable nonlinear Klein–Gordon equations and Toda lattices*, Comm. Math. Phys., 77 (1980), pp. 21–30.

SIAM J. MATH. ANAL.
Vol. 17, No. 4, July 1986

005

ON THE INFINITELY MANY SOLUTIONS OF A SEMILINEAR ELLIPTIC EQUATION*

C. JONES† AND T. KÜPPER‡

Abstract. A dynamical systems approach is developed for studying the spherically symmetric solutions of $\Delta u+f(u)=0$, where $f(u)$ grows like $|u|^{\sigma}u$ as $|u|\to\infty$. Various scalings are introduced to elucidate the singular behavior near the center and at infinity. The solutions of interest appear as trajectories in a three-dimensional phase space with a different amount of oscillation around a certain invariant axis. Using this oscillation, solutions with a prescribed number of zeros can be found when $\sigma<4/(n-2)$.

Key words. spherically symmetric solution, oscillation properties, dynamical systems

AMS(MOS) subject classifications. Primary 35P30, 35P05, 34B15, 34F15

1. Introduction. Variational arguments have been very successful in finding solutions of semilinear elliptic equations. Strauss [16] proved, by a nonlinear minimax argument, that a certain class of equations on $\mathbb{R}^n$ have infinitely many solutions. Berestycki and Lions [2] proved the same type of result allowing more general nonlinearities. A canonical example is given by the equation

$$\Delta_x u+|u|^{\sigma}u+\lambda u=0, \tag{1.1}$$

where Δ_x = Laplacian in $x\in R^n (n>1)$, $\lambda<0$, $\sigma>0$ and one is interested in classical L^2 solutions. Any of the above-mentioned results applied to this example guarantees the existence of infinitely many spherically symmetric solutions if $\sigma<4/(n-2)$.

Equation (1.1) is the standing wave equation for many nonlinear evolution equations, for instance the nonlinear Schrödinger, heat and wave equations. In nonlinear optics, this Schrödinger equation arises as a simplification to the Maxwell–Bloch equations. The solutions described in this paper may be relevant to some recent work on the behavior of an optical ring cavity, see McLaughlin, Moloney and Newell [11].

It has long been suspected that these infinitely many solutions are ordered, in some sense, by the number of zeros of the solution in the radial variable. Nehari [13] and Ryder [15] proved results that cover the case $n=3$, but the general case has remained open. The variational proofs mentioned above give only information about the energy values of the solutions and tell us nothing about the shape of their graphs.

In this paper we take a dynamical systems approach so as to realize the number of zeros as a geometrical property in phase space. Using this, we are able to prove for a large class of problems the existence of solutions with a prescribed number of zeros.

Spherically symmetric solutions of the equation

$$\Delta u+f(u)=0 \tag{1.2}$$

*Received by the editors May 17, 1984, and in revised form March 21, 1985.

†Department of Mathematics, University of Arizona, Tucson, Arizona 85721. Present address, Department of Mathematics, University of Maryland, College Park, Maryland 20742. Supported, in part, by the National Science Foundation under grants MCS 8200392, INT8314095 and by the Air Force Office of Scientific Research grant AFOSR 83 0227.

‡Abteilung der Mathematik, Universität Dortmund, 4600 Dortmund 50, West Germany. Supported in part by a grant from the Deutsche Forschungsgemeinschaft.

satisfy the ordinary differential equation

$$u_{rr}+\frac{n-1}{r}u_r+f(u)=0, \tag{1.3}$$

where r is the radial variable. We shall look for solutions of (1.3) that satisfy the boundary conditions

$$u_r(0)=0, \tag{1.4}$$

$$\lim_{r\to+\infty} u(r)=0. \tag{1.5}$$

Condition (1.4) is necessary to have a regular solution of (1.2), while (1.5) is forced (it turns out) by looking for an L^2 solution.

We make the following assumptions about $f(u)$:

(1) $f:\mathbb{R}\to\mathbb{R}$ is C^1.

(2) $f(u)=k(u)|u|^\sigma u+g(u)$, where

$$k(u)=\begin{cases} k_+ & \text{if } u\geqq 0,\\ k_- & \text{if } u<0,\end{cases} \quad \text{and } k_+>0,\ k_->0,$$

$$g(u)=0\big(|u|^\gamma\big),\quad g'(u)=0\big(|u|^{\gamma-1}\big) \quad \text{as } |u|\to+\infty, \quad \text{for some } \gamma<\sigma+1.$$

(3) $f(0)=0$, $f'(0)<0$ and if u is the smallest positive number for which $\int_0^u f(s)\,ds=0$, then u is not a critical point; similarly for the largest negative number for which

$$\int_0^u f(s)\,ds=0.$$

Remarks. (1) The function $f(u)=|u|^\sigma u+\lambda u$ with $\lambda<0$ and $\sigma>0$ is easily seen to satisfy these hypotheses.

(2) It is convenient to write $f(u)=k_\pm|u|^\sigma u+g(u)$, with the understanding that $k_\pm$ is k_+ if $u\geqq 0$ and k_- if $u<0$.

(3) Condition (3) seems a little strange and it is rather annoying that we need it. It says that in the phase portrait when $n=1$, 0 has two homoclinic orbits, one in the right half plane and the other in the left. This condition can be weakened somewhat as will be remarked in §4.

This paper is devoted to the proof of the following theorem.

THEOREM. *If $n>1$ and $\sigma<4/(n-2)$ then given m there is a solution of* (1.3)–(1.5), *$u(r)$ on $[0,\infty)$ with exactly m zeros in $[0,\infty)$. If $n=2$ these solutions exist for any σ.*

The relationship between σ and the space dimension n is the most subtle aspect of this problem. The Pohozaev [14] identity tells us that we cannot expect solutions to exist if $\sigma>4/(n-2)$. The variational proofs use the condition $\sigma<4/(n-2)$ for the compactness of a certain operator. It is interesting to see how the condition enters in the dynamical systems arguments.

Another drawback of the variational argument is that it needs the nonlinear term to be odd. This is not needed in our proof.

In §2 the basic framework in which we work will be given and explained. The basic idea of how the oscillation (number of zeros) is measured using a winding number in the plane is given in §3. This winding number idea is then used in §4 to obtain curves for the problem near $+\infty$ that have the necessary oscillation. Using some simple estimates, this is sufficient to obtain the solutions in the case $\sigma<2/(n-2)$ or $n=2$. This is done in §5. The full result is proved in §6; it needs another transformation and a much deeper understanding of the behavior near $r=0$.

2. Basic framework. Firstly, convert (1.3) to a system

$$u' = v, \qquad v' = -\frac{n-1}{r}v - f(u), \qquad ' = \frac{d}{dr}. \tag{2.1}$$

As $r \to +\infty$, (2.1) looks like the system when $n = 1$. We need to use this information in a concrete way. To do this, introduce the transformation

$$\rho = \frac{r}{r+1}. \tag{2.2}$$

Equation (2.1) then becomes

$$\begin{aligned} u' &= v, \\ v' &= -\frac{(n-1)(1-\rho)}{\rho}v - f(u), \qquad ' = \frac{d}{dr}. \\ \rho' &= (1-\rho)^2, \end{aligned} \tag{2.3}$$

Notice that the independent variable is still r, but r as a dependent variable has been changed. The plane $\rho = 1$ is now invariant ($\rho' = 0$) and carries the flow of (2.1) with $n = 1$. Equation (2.3) is, however, still singular at $\rho = 0$. This can be corrected by a change of independent variable which has the effect of multiplying the equation by ρ, call the new independent variable t

$$\begin{aligned} u' &= \rho v, \\ v' &= -(n-1)(1-\rho)v - \rho f(u), \qquad ' = \frac{d}{dt}. \\ \rho' &= \rho(1-\rho)^2, \end{aligned} \tag{2.4}$$

The plane $\rho = 1$ is still invariant and carries the flow associated to (2.1) with $n = 1$. The phase portrait for the example of $f(u) = |u|^a u + \lambda u$ is shown in Fig. 1.

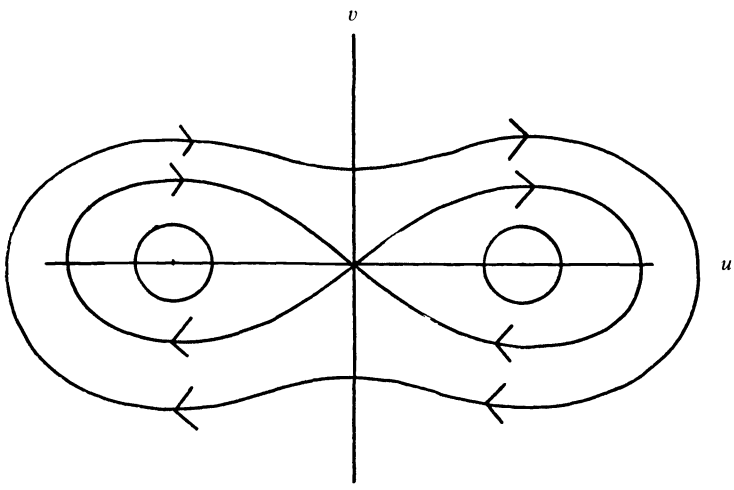


FIG. 1

From the assumption about $f(u)$, 0 must be a saddle and have two homoclinic orbits as shown in Fig. 2. However there may be more than one critical point inside each loop and there may be other critical points outside.

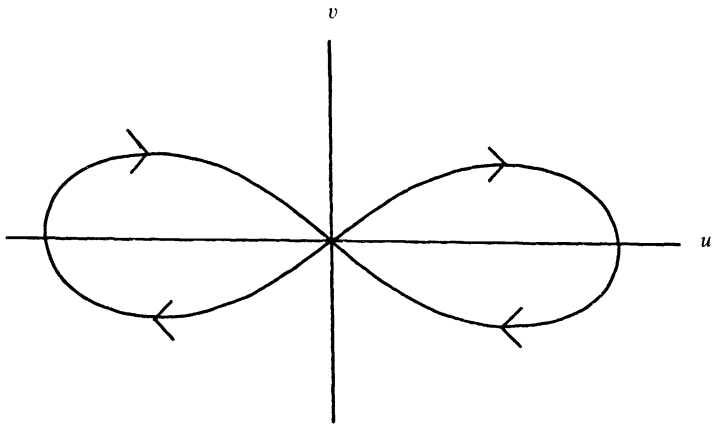


FIG. 2

If $\rho = 0$, the flow is that of the equation

$$u' = 0, \qquad v' = -(n-1)v, \tag{2.5}$$

for which the u-axis is a line of stable critical points. Notice that each vertical line is invariant. The flow is shown in Fig. 3.

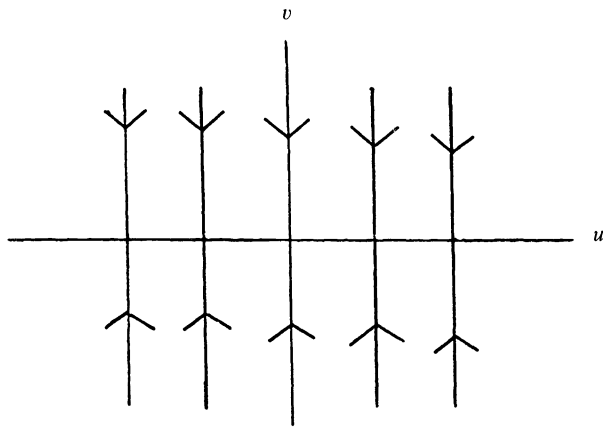


FIG. 3

The natural phase space for the problem (2.4) is $\mathbb{R}^2 \times [0,1]$. The plane $\rho = 0$ corresponds to $r = 0$ and $\rho = 1$ to $r = +\infty$. So the boundary conditions (1.4), (1.5) translate to looking for an orbit $(u(s), v(s), \rho(s))$ that satisfies

$$\lim_{s \to -\infty} (u(s), v(s), \rho(s)) = (a, 0, 0) \tag{2.6}$$

for some $a \neq 0$, and

$$\lim_{s \to +\infty} (u(s), v(s), \rho(s)) = (0,0,1). \tag{2.7}$$

In the full phase space such an orbit connects a critical point on the u-axis in $\rho = 0$ to the point $(0,0)$ in $\rho = 1$. The number of zeros is the number of times that the orbit crosses the plane $u = 0$. From the direction of the vector field on this plane, it can be seen that this corresponds to the number of oscillations around the invariant line $u = v = 0$. Therefore we are looking for solutions of the kind depicted in Fig. 4, where the amount of oscillation is prescribed.

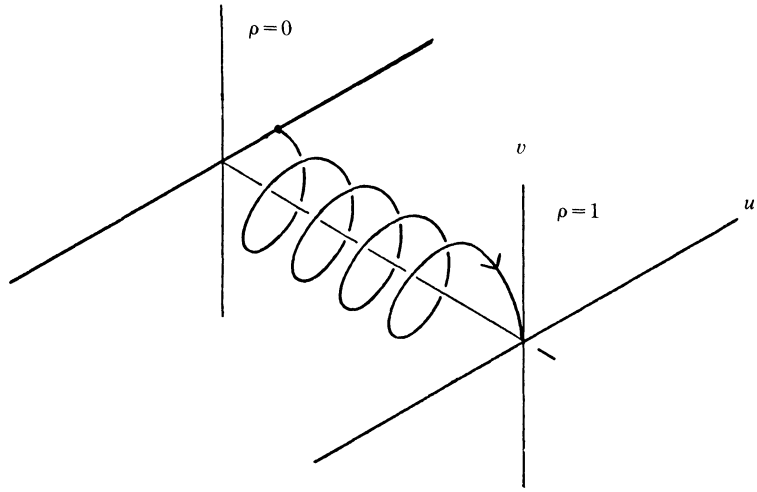


FIG. 4

The first step towards finding an orbit of the kind described above is to analyse the local behavior near the critical points to be connected.

The linearisation of (2.4) at $(0,0,1)$ is

$$\begin{pmatrix} 0 & 1 & 0 \\ -f'(0) & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \tag{2.8}$$

whose eigenvalues are 0, $\pm(-f'(0))^{1/2}$. The associated eigenvectors are $(0,0,1)$, $(1, \pm(-f'(0))^{1/2}, 0)$. These are the eigenvalues and eigenvectors of the one-dimensional system (Fig. 2) and a neutral eigenvalue whose eigenvector points away from the plane $\rho = 1$.

We can construct the center-stable manifold W^{cs}_{loc} at $(0,0,1)$. It is a small two-dimensional surface which contains a piece of the stable manifold of $(0,0)$ in $\rho = 1$, see Fig. 5.

It is easy to check that in a sufficiently small neighborhood of $(0,0,1)$, if $x = (u,v,\rho) \in W^{cs}_{loc}$ then $x \cdot t$ (the solution of (2.4) starting at x) tends to $(0,0,1)$ as $t \to +\infty$.

Therefore W^{cs}_{loc} contains points whose orbits satisfy the correct boundary condition (2.7).

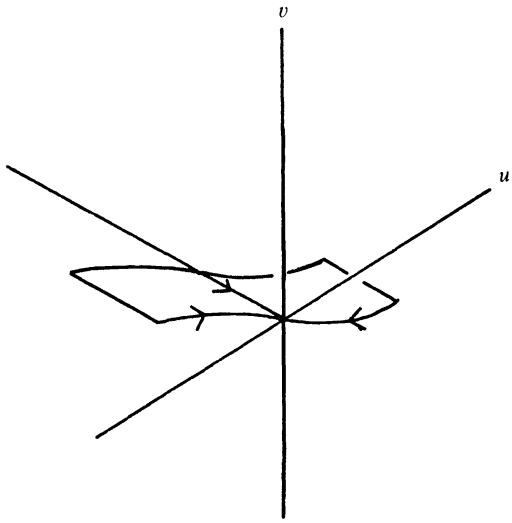


FIG. 5

At a critical point in the $\rho = 0$ plane, the linearisation is

$$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 1-n & f(a) \\ 0 & 0 & 1 \end{pmatrix}. \tag{2.9}$$

The eigenvalues of this matrix are 0, $1-n$, 1. The stable eigenvector is $(0,1,0)$. The eigenvectors associated to 0 and 1 are $(1,0,0)$ and $(0, -f(a)/n, 1)$ respectively. The unstable manifold points into the $\rho > 0$ space. This unstable manifold (call it $W^{u}_{loc}(a)$) corresponds to the unique solution of (1.3) that satisfies $u(0) = a$. This could actually be used as a proof of the uniqueness.

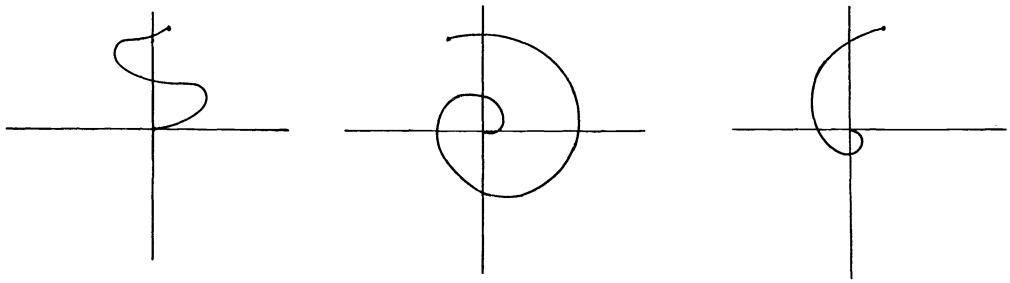


FIG. 6

Near a, these unstable manifolds can be collected together in the center-unstable manifold, W^{cu}_{loc}, see Fig. 6. These manifolds are unique because they are negatively invariant; see Jones [9] for this type of argument. The idea here is that if there were two

surfaces with center-unstable behavior, the complementary stable direction would tear them apart (in backward time).

Now set $W^0=\bigcup_{\alpha\in\mathbb{R}} W^u_{\text{loc}}(a)$. W^0 is then a surface containing the line $\{(u,v,\rho): v=\rho=0\}$. It contains all the solutions to the original problem which are regular at $r=0$.

The argument will proceed by iterating W^{cs}_{loc} of $(0,0,1)$ in backward time and making it intersect W^0. Part of the problem is that we cannot get an estimate on how far W^0 extends in the ρ direction that is uniform in a.

3. Winding number. In this section we define a winding number for a certain type of curve in $\mathbb{R}^2$. It will then be shown how we use this to measure the oscillation of solutions.

Let C be a curve in $\mathbb{R}^2$, that is C is given by a function $\phi:[s_0,s_1]\to\mathbb{R}^2$.

DEFINITION. C is called an admissible curve if the following are satisfied:

(1) ϕ is continuous and $\phi(s)\neq 0$ for all $s\in(s_0,s_1]$.

(2) $\phi(s_0)=0$ and a tangent vector to C exists at s_0.

This last condition says that

$$\lim_{s\to s_0}\frac{\phi(s)}{s-s_0}$$

exists and is nonzero; we shall always denote it ψ_0.

The main point about an admissible curve is that we can assign an angle to each point on it, $\theta(x)$ according to the rules:

(1) $\theta(\phi(s_0))=\arg(\psi_0)$ (argument of ψ_0), where $\arg(\psi_0)\in[0,2\pi)$.

(2) $\theta(\phi(s))$ is continuous in $s\in[s_0,s_1]$.

This is possible from the definition of admissibility. If $\phi(s)$ is written as $r(s)e^{i\theta(s)}$, where a suitable branch for $\theta(s)$ is chosen, then

$$r(s)e^{i\theta(s)}/(s-s_0)\to\psi_0$$

as $s\to s_0$. If $\psi_0=r_0e^{i\theta_0}$, we have $r(s)/(s-s_0)\to r_0$ and $\theta(s)\to\theta_0$.

DEFINITION. The winding number of an admissible curve C is then defined to be

$$I(C)=\left[\frac{1}{2}\left(\frac{2\theta(\phi(s_1))}{\pi}+1\right)\right]-\left[\frac{1}{2}\left(\frac{2\theta(\phi(s_0))}{\pi}+1\right)\right]$$

where $[x]=$ greatest integer less than or equal to x.

This rather strange looking definition becomes less mysterious when it is noticed that the function

$$\frac{1}{2}\left(\frac{2x}{\pi}+1\right)$$

maps the points $-\pi/2$, $\pi/2$, $3\pi/2,\cdots$ into the points $0,1,2,\cdots$, respectively. Thus this winding number counts the net number of crossings of the vertical axis in $\mathbb{R}^2$. For example, the curves in Fig. 6 have winding numbers 0, 3 and -2.

This winding number does not measure the exact number of crossings of the vertical axis by these curves. It seems then that this number could only be used to obtain a lower bound on the number of zeros and not give the exact number. However, it turns out that the winding of the relevant curves gives the exact number of zeros. The

curve used is not the solution but a curve on W^{cs}. It is the power of this method that the topological winding of this curve determines the exact number of zeros of the solution with initial value at the end point of this curve. See Proposition 3.5 below.

It is convenient to give a covering space interpretation to this winding number as then the action of the flow as a homotopy can be viewed easily. The problem of dealing with curves that go to 0 but have well defined tangent vectors there can be dealt with by blowing up the origin and then using the covering space of this space.

Consider polar co-ordinates on $\mathbb{R}^2\setminus\{0\}$. This is a mapping of $\mathbb{R}^2\setminus\{0\}$ to $\mathbb{R}^+\times S^1$, $(s,\theta)\in\mathbb{R}^+\times S^1$ and $\mathbb{R}^+=\{r\in\mathbb{R}: r>0\}$. $\mathbb{R}^+\times S^1$ can be completed to $\mathbb{R}^+\times S^1$ by adding on the set $r=0$. Let $j:\mathbb{R}^2\setminus\{0\}\to\mathbb{R}^+\times S^1$ be the inclusion given by polar co-ordinates. Set $X=\mathbb{R}^+\times S^1$.

Let C be an admissible curve in $\mathbb{R}^2$. We claim that a curve $\hat{C}$ in X can be naturally associated to C as follows. If C is given by $\phi:[s_0,s_1]\to\mathbb{R}^2$, $\hat{C}$ will be given by $\hat{\phi}:[s_0,s_1]\to X$. For $s\in(s_0,s_1]$, define

$$\hat{\phi}(s)=j(\phi(s)).$$

For $s=s_0$, set $\hat{\phi}(s_0)=(0,\arg\psi_0)$, where ψ_0 is the tangent vector to C at s_0. To check this is continuous; for s near s_0, write $\phi(s)=r(s)\exp\{i\theta(s)\}$. Then

$$j(\phi(s))=(r(s),\theta(s))\to\Big(0,\lim_{s\to s_0}\theta(s)\Big).$$

But

$$\lim_{s\to s_0}\phi(s)/(s-s_0)=\psi_0$$

and

$$\phi(s)/(s-s_0)=\{r(s)/(s-s_0)\}\exp\{i\theta(s)\}.$$

This implies that if $\psi_0=\rho\exp\{i\gamma\}$, then

$$\theta(s)\to\gamma \quad \text{as } s\to s_0.$$

The space $X=\mathbb{R}^+\times S^1$ admits $\tilde{X}=\mathbb{R}^+\times\mathbb{R}$ as a covering space. Since we wish to measure crossings of the v-axis, it is convenient to use a covering map that sends vertical lines at integer intervals to the lines $\theta=\pi/2,3\pi/2,\cdots$. If $(r,x)\in\tilde{X}$, such a map is:

$$p(r,x)=\left(r,2\pi\left(\frac{2x-1}{4}-\left[\frac{2x-1}{4}\right]\right)\right). \tag{3.1}$$

The line $x=0$ is sent to $\theta=3\pi/2$ and $x=1$ to $\theta=\pi/2$, etc.

From standard covering space theory, see Munkres [12], the curve $\hat{C}$ can be lifted to a curve $\tilde{C}$ in $\tilde{X}$ and this is unique given the choice of starting point. For each $q\in p^{-1}(\phi(s_0))$, there is a unique curve $\tilde{C}$ so that $p(\tilde{C})=\hat{C}$.

The winding number of C, $I(C)$, is now determined by the number of crossings of lines $x=0,\pm1,\pm2,\cdots$ that $\tilde{C}$ makes. If $\tilde{C}$ is given by $\tilde{\phi}:[s_0,s_1]\to\tilde{X}$ and

$$\tilde{\phi}(s)=(r(s),x(s)),$$

then this is

$$I(C)=[\tilde{x}(s_1)]-[\tilde{x}(s_0)].$$

It is easy to check that this is independent of the choice of $\tilde{\phi}(s_0)$.

Now consider how maps on $\mathbb{R}^2$ lift to this covering space. This will lead to lifting homotopies.

Let $\Phi:\mathbb{R}^2\to\mathbb{R}^2$ be continuous and satisfy:

(1) $\{0\}$ and $\mathbb{R}^2\setminus\{0\}$ are preserved;

(2) Φ is differentiable at $x=0$ and $D\Phi(0)$ is nonsingular.

There is naturally associated to such a Φ a map $\hat{\Phi}:X\to X$ so that

$$\hat{\Phi}(j(x))=j(\Phi(x))$$

for $x\in\mathbb{R}^2\setminus\{0\}$. Recall that $j:\mathbb{R}^2\setminus\{0\}\to\mathbb{R}^+\times S^1=X^0$ is the polar co-ordinates map. It is trivial to define Φ^0 on X^0 to commute with j,

$$\Phi(r,\theta)=(R(r,\theta),\gamma(r,\theta))$$

where

$$\Phi(re^{i\theta})=R(r,\theta)\exp\{i\gamma(r,\theta)\}.$$

Since Φ is differentiable at 0

$$\lim_{r\to 0}\frac{1}{r}\Phi(re^{i\theta})=D\Phi(0)e^{i\theta}.$$

Since $D\Phi(0)$ is nonsingular

$$\lim_{r\to 0}\frac{R(r,\theta)}{r}$$

is nonzero and

$$\lim_{r\to 0}\gamma(r,0)=\arg\{D\Phi(0)e^{i\theta}\}. \tag{3.2}$$

It follows that Φ^0 can be extended to X by setting

$$\Phi(0,\theta)=\arg\{D\Phi(0)e^{i\theta}\}.$$

Φ is thus continuous in r and θ separately. Further $R(r,\theta)\to 0$ uniformly in θ since Φ is continuous at 0 and (3.2) is uniform in θ from the definition of differentiability. It follows that Φ is continuous on X.

Now consider the case of a homotopy $\Phi:\mathbb{R}^2\times[a,b]\to\mathbb{R}^2$ under the following assumptions:

(1) $\{0\}$ and $\mathbb{R}^2\setminus\{0\}$ are preserved by Φ_t for each t. ($\Phi_t(x)=\Phi(x,t)$).

(2) Φ_t is differentiable at $x=0$, for $t\in[a,b]$. $D_x\Phi_t(0)$ is nonsingular and continuous in t.

(3) $|\Phi_t(x+h)-\Phi_t(x)-D_x\Phi_t(0)h|\to 0$ uniformly in $t\in[a,b]$.

DEFINITION. A homotopy $\Phi:\mathbb{R}^2\times[a,b]\to\mathbb{R}^2$ is called admissible if (1)–(3) above are satisfied.

Let $\hat{j}(\mathbb{R}^2\setminus\{0\})\times[0,1]\to X\times[0,1]$ be $j\times id$.

PROPOSITION 3.1. *If Φ is an admissible homotopy, then there exists a homotopy $\hat{\Phi}:X\times[0,1]\to X$ such that*

$$\hat{\Phi}\circ\hat{j}=j\circ\Phi \tag{3.3}$$

on $(\mathbb{R}^2\setminus\{0\})\times[0,1]$.

Proof. The above considerations about a map on $\mathbb{R}^2$ guarantee that Φ_t is defined for each t and (3.3) is satisfied. It remains to show that $\hat{\Phi}$ is continuous. From the above reasoning $\hat{\Phi}_t$ is continuous in (r,θ). Since $D_x\Phi_t(0)$ is continuous in t, $\hat{\Phi}_t$ is continuous in t for fixed (r,θ). Joint continuity then follows from (3).

The space $X=\mathbb{R}^+\times S^1$ is covered by $\tilde{X}=\mathbb{R}^+\times\mathbb{R}$ as mentioned earlier. From standard covering space theory, $\hat{\Phi}$ lifts to a homotopy $\tilde{\Phi}$ on $\tilde{X}$. So

$$\tilde{\Phi}:\tilde{X}\times[0,1]\to\tilde{X}.$$

Let C be an admissible curve and Φ an admissible homotopy. It is obvious that $\Phi_t(C)$ is also an admissible curve, for any $t\in[a,b]$. Since C is not a closed curve, $I(\Phi_t(C))$ can change with t. This is measured in $\tilde{X}$ by $\tilde{\Phi}_t(\tilde{C})$.

Let $A\subset\tilde{X}$ be given by

$$A=\{(r,x):x\in\mathbb{Z}\}.$$

The proof of the following proposition is immediate from the winding number.

PROPOSITION 3.2. *If* $\tilde{\Phi}_t(\tilde{\phi}(s_0))\in A$ *and* $\tilde{\Phi}_t(\tilde{\phi}(s_1))\in A$ *for all* $t\in[a,\tau]$ *then*

$$I(\Phi_t(C))=I(\Phi_0(C)).$$

The next task is to relate this number and homotopy to the flow of (2.4).

Let $w=(u,v,\rho)\in\mathbb{R}^2\times[0,1]$, in the notation of (2.4). Rewrite (2.4) as

$$w'=F(w),\qquad '=\frac{d}{dt}. \tag{3.4}$$

From the assumption that $f(u)$ is C^1, $F(w)$ is easily seen to be C^1 on $\mathbb{R}^2\times(0,1)$. Let $\psi(x,t)$ be the flow operator for (3.4). Now restrict x to lie in the set $\{\rho=\gamma\}\simeq\mathbb{R}^2$ for some $\gamma\in(0,1)$.

We claim that $\psi(x,t)$ is defined for all $t\in\mathbb{R}$. For $t>0$, this follows from the fact that energy is decreasing for (2.1). Set

$$H(u,v)=\frac{v^2}{2}+\int_0^u f(s)\,ds; \tag{3.5}$$

then along orbits of (2.1)

$$\frac{dH}{dr}=-\frac{(n-1)}{r}v^2.$$

Consequently in forward r, the solution to (2.1) with $(u(r_0),v(r_0))=(u_0,v_0)$, is constrained to lie inside $H(u,v)=H(u_0,v_0)$ and so exist for all $r\geqq 0$. This property then transfers easily to (2.3) and (2.4). That solutions to (3.4) exist for $t<0$ follows from Corollary 5.1 to be proved in §5. This gives an a priori estimate to solutions of (2.1) with $r\leqq r_0$.

Consider Ψ as a map on $\{\rho=\gamma\}\times[a,b]$ some $a,b\in\mathbb{R}$. From the above it is defined and continuous with values in $\mathbb{R}^2\times[0,1]$. Let $\pi:\mathbb{R}^2\times[0,1]\to\mathbb{R}^2$ be the natural projection and set $\Phi=\pi\circ\Psi$. Now consider Φ as a homotopy on $\mathbb{R}^2$,

$$\Phi:\mathbb{R}^2\times[0,1]\to\mathbb{R}^2 \tag{3.6}$$

with the understanding that the domain $\mathbb{R}^2$ depends on γ.

We must check that Φ is an admissible homotopy. Condition (1) is satisfied because $u=v=0$ is a solution of the equation. Since Φ is a restriction of the solution operator of (3.4), (2) follows because F is C^1. (3) is the only condition that is not

immediate. But the fact that the difference quotient of the solution operator approaches the derivative uniformly on compact intervals of t is part of the proof that it converges. In particular it follows from the proof of Theorem 3.1 in Hartman [8, p. 98]. (3) would follow easily from assuming that f is C^2 but in our main example

$$f(u)=|u|^{\sigma}u+\lambda u,$$

f may only be C^1 at $u=0$, if $0<\sigma<1$. We have proved the following proposition.

PROPOSITION 3.3. *Φ is an admissible homotopy.*

Recall W^{cs}_{loc} from §2; this is the local center-stable manifold of $(0,0,1)$ for (2.4). It consists of orbits that satisfy the desired boundary condition at $+\infty$. Let

$$W^{cs}=\bigcup_{t\leqq 0} W^{cs}_{loc}\cdot t=\bigcup_{t\leqq 0}\Psi(W^{cs}_{loc},t).$$

W^{cs} is the global center-stable manifold. The main curves whose winding number will be measured are ones that lie in $W^{cs}\cap\{\rho=\gamma\}$.

Let $x\in W^{cs}\cap\{\rho=\gamma\}$ and C_x be a compact curve in $W^{cs}\cap\{\rho=\gamma\}$ joining x to $(0,0)$. In other words, if C_x is given by $\phi:[s_0,s_1]\to\{\rho=\gamma\}$, $\phi(s_0)=0$ and $\phi(s_1)=x$. We claim that such a curve C_x is admissible.

Because the manifold W^{cs}_{loc} for (2.4) is unique, it contains a piece of the curve $u=v=0$ for ρ near 1. This line separates W^{cs}_{loc} into two pieces

$$W^{cs}_{loc}=W^R\cup W^L$$

where $W^R\subset\{u>0\}$ and $W^L\subset\{u<0\}$. The point $x\in W^{cs}_{loc}\cdot t$ some t, and therefore it lies in $W^R\cdot t$ or $W^L\cdot t$. Assume without loss of generality that $x\in W^R\cdot t$ and that $C_x\setminus\{\phi(s_0)\}\subset W^R\cdot t$.

Now ϕ is clearly continuous and $\phi(s)\neq 0$ for $s\neq s_0$. We only have to show that a tangent vector exists at s_0. $C_x\cdot t\subset W^R$ and $\phi(s_0)\cdot t\in\{(0,0)\}\times[0,1]$. Since ρ' is independent of u and v in (2.4) $(C_x\setminus\{\phi(s_0)\})\cdot t\subset W^R\cap\{\rho=\bar\gamma\}$ some $\bar\gamma$. Since W^{cs}_{loc} is smooth and transverse to $\{\rho=\bar\gamma\}$, $W^R\cap\{\rho=\bar\gamma\}$ is a one-dimensional curve in $\{\rho=\bar\gamma\}$ and has a tangent vector at $(0,0)$. It therefore follows that $C_x\cdot t$ does also. It is now easy to see that C_x has a tangent vector at s_0 and we have proved the following proposition.

PROPOSITION 3.4. *If C_x is a continuous, compact curve in $W^{cs}\cap\{\pi=\gamma\}$ given by $\phi:[s_0,s_1]\to\{\rho=\gamma\}$ with $\phi(s_0)=0$ and $\phi(s_1)=x\neq 0$, then C_x is admissible.*

If $x\in W^{cs}$, the solution to (2.4), $x(t)$ with $x(0)=x$ satisfies $x(t)\to(0,0,1)$ as $t\to+\infty$. The reason our approach works is the following proposition.

PROPOSITION 3.5. *If C_x is a curve as described in Proposition* 3.4, *then*

$$I(C_x)=\text{number of zeros of } u(t) \text{ in the solution } x(t),\ t\in[0,\infty).$$

Remark. This proposition is not obvious because C_x is not the solution curve, it is a curve in W^{cs}.

Proof. $C_x\subset W^{cs}\cap\{\rho=\gamma\}$, we can apply the homotopy Φ derived from the flow. If τ is a large positive number, then

$$I(\Phi_\tau(C_x))=0 \tag{3.7}$$

since $\Phi_\tau(C_x)\subset W^R$ (or W^L). From Proposition 3.2 there exists $t\in[0,\tau]$ with $\tilde\Phi_t(\tilde\phi(s_0))$ or $\tilde\Phi_t(\tilde\phi(s_1))$ in the set A.

The tangent vector to $W^R \cap \{\rho = \bar{\gamma}\} \cdot t$, for all $t < 0$, lies in the set $\{u > 0\}$. This can be seen by checking the vector field on $u = 0$ and $H(u,v) = 0$ for (2.3), see Fig. 7. Notice that it points for large $\bar{\gamma}$ and $t = 0$ into the region between the bulb $H(u,v) = 0$ and $\{u = 0,\ v < 0\}$.

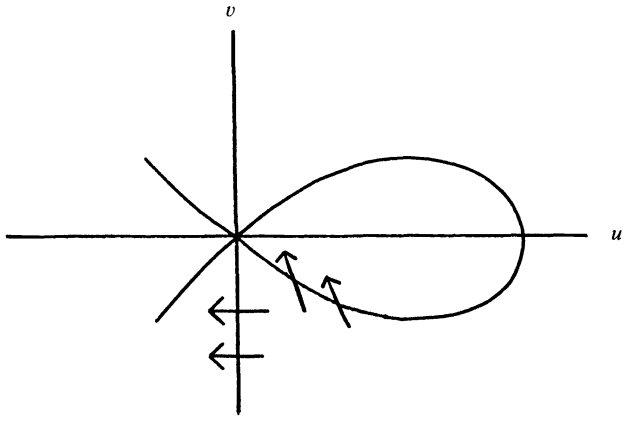


FIG. 7

It follows that $\hat{\Phi}_t(\hat{\phi}(s_0))$ never crosses $\theta = \pi/2$ or $\theta = 3\pi/2$. Consequently $\tilde{\Phi}_t(\tilde{\phi}(s_0)) \in A$ for all $t \in [0,\tau]$.

Therefore $\tilde{\Phi}_t(\tilde{\phi}(s_1)) \in A$ for some $t \in [0,\tau]$. In fact it must cross A at least $I(C_x)$ times for $t \in [0,\tau]$, from (3.7).

Now $\tilde{\Phi}_t(\tilde{\phi}(s_1)) \in \tilde{X} \times [0,1]$ and $p(\tilde{\Phi}_t(\tilde{\phi}(s_1))) = \Phi_t(\phi(s_1)) = \Phi_t(x) = \pi(\psi_t(x))$. But $\psi_t(x) = w(t)$ is the solution of (2.4) satisfying $w(0) = x$. If $\tilde{\Phi}_t(\tilde{\phi}(s_1)) \in A$ then $\Phi_t(x) \in \{u = 0\}$, which means that if $w(t) = (u(t), v(t), \rho(t))$ u has a zero at that value of t. The direction of the vector field of (2.4) on the v-axis is determined by $u' = \rho v$. Since we can assume $\rho > 0$, it is as given in Fig. 8.

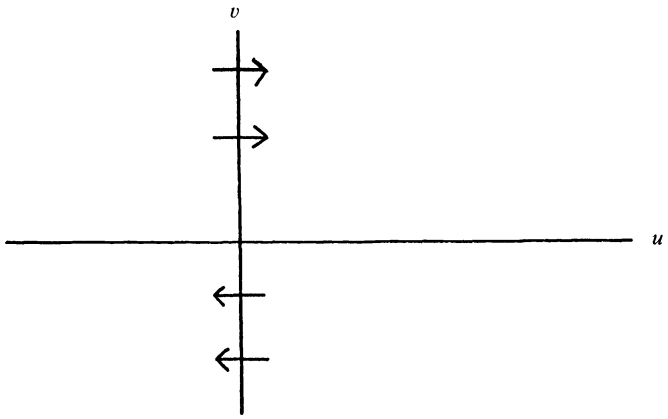


FIG. 8

A solution can only cross $v=0$ with angle decreasing in forward time. In the covering space, $\tilde{\Phi}_t(\tilde{\phi}(s_1))$ can only cross a line A with decreasing horizontal co-ordinate. It then follows that it must cross exactly $I(C_x)$ times. Therefore $u(t)$, $t\in[0,\infty)$, has $I(C_x)$ zeros.

As a consequence of the proof above we can see that $I(\Phi_t(C_x))$ can only increase as t decreases with $t<0$.

PROPOSITION 3.6. *If C_x is a curve as described in Proposition* 3.4, *$I(\Phi_t(C_x))$ is a decreasing function of t, for all $t\in\mathbb{R}$.*

4. Behavior near $+\infty$. The proof of the theorem will proceed by iterating W^{cs}_{loc} of $(0,0,1)$, in backward time and ensuring that it eventually intersects W^0. The zeros of the solution will be obtained by constructing curves in W^{cs} with a certain amount of winding. It was proved in §3 that the winding of such a curve determines exactly the number of zeros of the solution starting at the end point. In this section we shall prove the existence of curves in W^{cs} with arbitrarily large winding number.

Recall the basic equation (2.4):

$$\begin{aligned} u' &= \rho v, \\ v' &= -(n-1)(1-\rho)v - \rho f(u), \\ \rho' &= \rho(1-\rho)^2. \end{aligned}$$

Let $W\subset\mathbb{R}^2\times[0,1]$ be given by

$$W=\{(u,v,\rho): 1-\gamma\leqq\rho\leqq 1\},$$

$\gamma>0$ is fixed.

We claim that W is a Wazewski set in backward time, see Conley [5] for definition. Let

$$\begin{aligned} W^0 &= \{x\in W: x\cdot t\in W \text{ some } t<0\}, \\ W^+ &= \{x\in W: x\cdot(t,0]\notin W \text{ for all } t<0\}. \end{aligned}$$

For W to be a Wazewski set, it suffices that W and W^+ be closed in $\mathbb{R}^2\times[0,1]$. Since $W^+=\{(u,v,\rho):\rho=1-\gamma\}$, this is obvious. The point about a Wazewski set is that one can define a continuous map $R: W^0\to W^+$ as follows. For $x\in W^0$, let

$$\tau(x)=\sup\{t<0: x\cdot t\notin W\},$$

and set $R(x)=x\cdot\tau(x)$. Wazewski's principle says that $R(x)$ is a retract of W^0 to W^+. It maps a point $x\in W^0$ to the place where it first leaves W.

Now choose γ so that $W^{cs}\cap\{\rho=1-\gamma\}$ is nonempty. In fact $W^{cs}\cap\{\rho=1-\gamma\}$ will be a curve containing $(0,0)$. Let Γ be a continuous curve in W^{cs} satisfying the conditions:

(1) $(0,0,1-\gamma)\in\Gamma$.

(2) $\Gamma\setminus\{(0,0,1-\gamma)\}\subset\{u>0\}$.

(3) There exists exactly one point $\bar{y}$ in $\Gamma\cap\{\rho=1\}$, see Fig. 9.

The set $\Gamma^0=\Gamma\setminus\{\bar{y}\}\subset W^0$ and so we can restrict R to Γ^0 as a continuous map; it is then obvious that $R(\Gamma^0)\subset W^{cs}$. The following lemma is what we need about the behavior near $\rho=1$ ($r=+\infty$) for the existence theorem.

LEMMA 4.1. *There exists $x\in R(\Gamma^0)\subset W^{cs}$ such that $C_x\subset R(\Gamma^0)$ and $I(C_x)$ is arbitrarily large.*

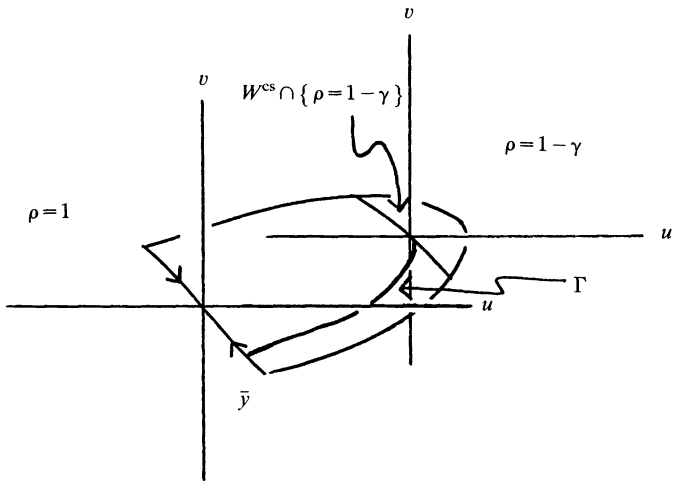


FIG. 9

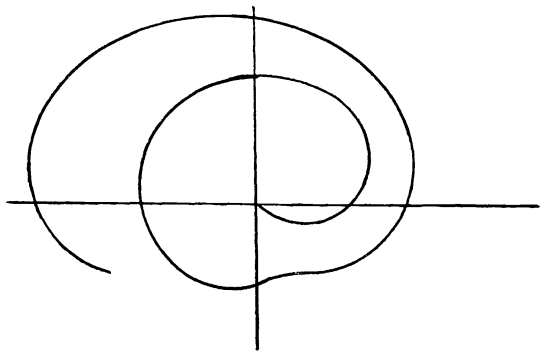


FIG. 10

The lemma says that $R(\Gamma^0)$ contains curves as shown in Fig. 10.

The idea is that because of energy dissipation, the solutions will move outside the "bow-tie" of $\{\rho = 1\}$ but the oscillation in $\{\rho = 1\}$ will cause them to rotate around $(0,0)$. There is, however, a nontrivial estimate to be made. Since the solutions are moving away from $\{\rho = 1\}$, it is a little delicate to check that the oscillation can still be forced on them.

Recall from §2 that $(0,0,1)$ has one negative, one positive and one zero eigenvalue. Let E^s, E^u and E^c be the associated eigenspaces, with (y,z,τ) as the co-ordinates given by the decomposition $E^s \oplus E^u \oplus E^c$. We can assume that $\tau = 1-\rho$. There is a local center-unstable manifold W^{cu}_{loc} given by

$$y = g_1(z,\tau)$$

in a neighborhood of $(0,0,0)$. W^{cs}_{loc} is given by $z = g_2(y,\tau)$. Now change co-ordinates on $\mathbb{R}^2 \times [0,1]$ so that near $(0,0,1)$, W^{cu}_{loc} and W^{cs}_{loc} become orthogonal planes. Near $(0,0,1)$

these could be:

$$\zeta = y - g_1(z, 1-\rho),$$
$$\eta = z - g_2(y, 1-\rho),$$
$$\tau = 1-\rho.$$

Near the origin the equations now take the form

$$\begin{aligned} \zeta' &= \alpha\zeta + f_1(\zeta,\eta,\tau), \\ \eta' &= \beta\eta + f_2(\zeta,\eta,\tau), \\ \tau' &= -(1-\tau)\tau^2 \end{aligned} \tag{4.1}$$

where $\alpha < 0$, $\beta > 0$ and $f_i(\zeta,\eta,\tau) = O(|\zeta|+|\eta|+|\tau|)$, $i = 1$ or 2. The relevant parts of the new phase portrait are depicted in Fig. 11.

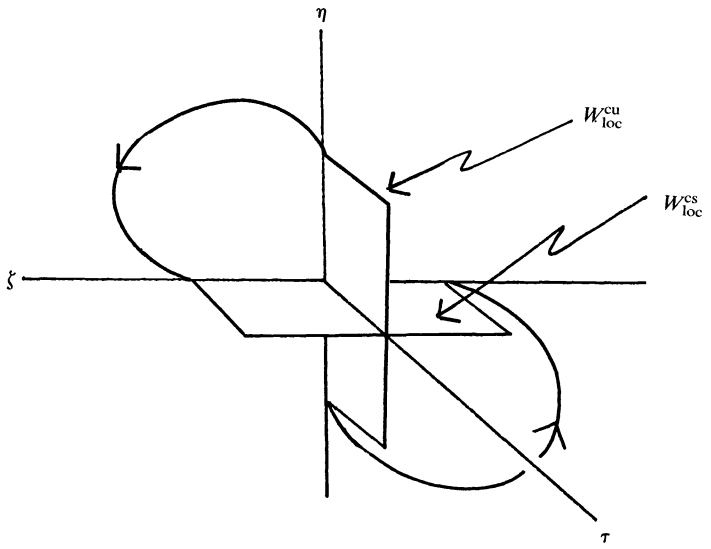


FIG. 11

Set $V = V(\varepsilon,\delta) = [-\varepsilon,\varepsilon]\times[-\varepsilon,\varepsilon]\times[0,\delta]$ in (ζ,η,τ) space. If ε is small enough, V is a backwards Wazewski set. We shall need some notation

$$z_1^+ = W^u_{loc} \cap \{\eta = \varepsilon\}, \qquad z_2^+ = W^u_{loc} \cap \{\eta = -\varepsilon\},$$
$$z_1^- = W^s_{loc} \cap \{\zeta = \varepsilon\}, \qquad z_2^- = W^s_{loc} \cap \{\zeta = -\varepsilon\}.$$

It is clear that these sets are just single points. Recall that $W^s_{loc} = W^{cs}_{loc} \cap \{\tau = 0\}$. Let D_1^+ be a neighborhood of z_1^+ in $\{\eta = \varepsilon\}$ and D_1^- a neighborhood of z_1^- in $\{\zeta = \varepsilon\}$. Define D_2^+ and D_2^- similarly, see Fig. 12.

Let $P: V^0 \to V^+$ be the Wazewski map for V, where V^0 and V^+ have the analogous meaning to W^0 and W^+. P is defined on $D_1^+ \setminus \{z_1^+\}$ and $D_2^+ \setminus \{z_2^+\}$. We can also define a map $Q: D_1^- \to \{\eta = \varepsilon\}$ as follows. Let $T(z_1^-)$ be determined by $z_1^- \cdot T(z_1^-) = z_1^+$. If $\pi(\zeta,\eta,\tau) = \eta$ and Φ is still the flow,

$$\pi(\Phi(z,t)) = \varepsilon$$

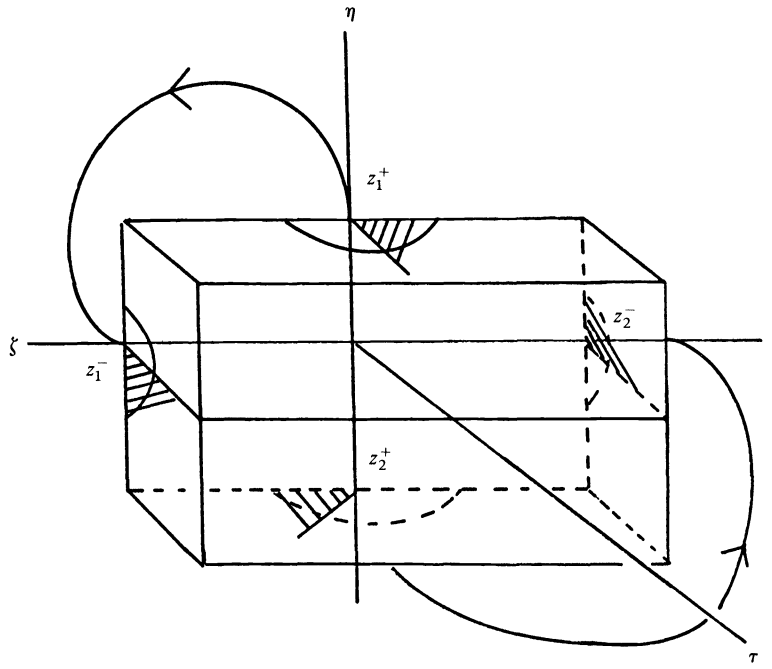


FIG. 12

can be solved, t as a function of z, near $(z_1^-, T(z_1^-))$. This follows from the implicit function theorem, knowing that the vector field is transverse to $\eta = \varepsilon$. Let the solution be denoted $T = T(z)$. Now restrict D_1 to lie inside the domain of $T(z)$ and set

$$Q_1(z) = z \cdot T(z).$$

Similarly define Q_2 on D_2^-.

We shall need an estimate on what Q_1, Q_2 do outside the box and one on what P does inside. Let $K_a^1 = \{(\zeta, \eta, \tau)\colon \eta = \varepsilon \text{ and } \zeta \leqq -a\tau\}$, $K_a^2 = \{(\zeta, \eta, \tau)\colon \eta = \varepsilon \text{ and } \zeta \geqq a\tau\}$, see Fig. 12.

LEMMA 4.2. *If D_1^- is small enough,*

$$Q_1(D_1^- \cap \{\eta \leqq 0\}) \subset K_a^1.$$

LEMMA 4.3. *If D_2^- is small enough,*

$$Q_2(D_2^- \cap \{\eta \geqq 0\}) \subset K_a^2.$$

LEMMA 4.4. *If D_1^+ is small enough, then any set $A \subset K_a^1 \cap D_1^+$ with $z_1^+ \notin A$ satisfies*

$$P(A) \subset \{\zeta = -\varepsilon\} \cap \{\eta \geqq 0\} \cap D_2^-$$

where D_2^- is some prescribed neighborhood of z_2^-.

LEMMA 4.5. *If D_2^+ is small enough, then any set $A \subset K_a^2 \cap D_2^+$ with $z_2^+ \notin A$ satisfies*

$$P(A) \subset \{\zeta = \varepsilon\} \cap \{\eta \leqq 0\} \cap D_1^-$$

where D_1^- is a prescribed neighborhood of z_1^-.

We shall only prove Lemmas 4.2 and 4.4; 4.3 and 4.5 only involve changing notation. We shall do 4.4 first.

Proof of Lemma 4.4. The set $\{\eta=\varepsilon\}\setminus\{z_1^+\}\subset V^0$, therefore P is defined on A. Since $\eta=0$ and $\zeta=0$ are invariant relative to V, it is clear that $P(A)\subset\{\eta\geqq 0\}\cap\{\zeta\leqq 0\}$. The lemma will follow from proving the following two facts; a is fixed and positive:

(1) If D_1^+ is small enough, $P(K_a^1\cap D_1^+)$ lies in a prescribed neighborhood of $\{\tau=0\}$.

(2) If $z_k\to z_1^+$ and $\{z_k\}\subset K_a^1$ then $P(z_k)\to z_1^+$.

Note that (2) does not trivially follow from the continuity of P because P is not defined at z_1^+. We firstly prove (1). This is the key estimate, for it says that the influence of the flow in $\{\tau=0\}$ is stronger than that of τ'.

Let $z=(\zeta_0,\eta_0,\tau_0)\in K_a$, some $a>0$. Define

$$T_\tau(z)=\inf\{|t|: t<0 \text{ and } z\cdot t\in\{\tau=\delta\}\},$$
$$T_\zeta(z)=\inf\{|t|: t<0 \text{ and } z\cdot t\in\{\zeta=-\varepsilon\}\}.$$

(1) follows from showing that

$$T_\zeta(z)/T_\tau(z)\to 0 \quad \text{as } z\to z_1^+, \tag{4.2}$$

for then the orbit $z\cdot t$ will reach $\zeta=-\varepsilon$ before it reaches $\tau=\delta$, and this will be true for any δ, making $|z-z_1^+|$ small enough.

From (4.1)

$$\zeta'=\alpha\zeta+f_1(\zeta,\eta,\tau), \tag{4.3}$$

with $\alpha<0$. On any orbit of interest $\zeta\leqq 0$. Since $\zeta=0$ is invariant,

$$f_1(\zeta,\eta,\tau)=\zeta g(\zeta,\eta,\tau)$$

where $g(\zeta,\eta,\tau)$ can be made as small as desired by making V small. Therefore

$$\frac{d\zeta}{dt}\geqq\frac{\alpha}{2}\zeta.$$

Integrating, we have

$$\int_{-\varepsilon}^{\zeta_0}\frac{d\zeta}{\alpha\zeta}\geqq\int_{-T_\zeta}^{0}\frac{1}{2}dt$$

so,

$$\frac{2}{\alpha}\ln\frac{\zeta_0}{-\varepsilon}\geqq T_\zeta.$$

This is the estimate on T_ζ. For τ we have from (4.1) in V, $\tau'\geqq-\tau^2$ and so

$$\int_\delta^{\tau_0}\frac{d\tau}{\tau^2}\geqq-\int_{T_\tau}^{0}dt$$

which gives

$$-\frac{1}{\tau_0}+\frac{1}{\delta}\geqq -T_\tau$$

or,

$$T_\tau\geqq\frac{1}{\tau_0}-\frac{1}{\delta}\geqq\frac{1}{2\tau_0},$$

if τ_0 is chosen so that $\delta > 2\tau_0$, for instance. But then

$$T_\zeta / T_\tau \leqq \frac{4\tau_0}{\alpha} \ln \left| \frac{\zeta_0}{\varepsilon} \right|.$$

Since $z \in K_a$, $|\zeta_0| \geqq a\tau_0$, and so

$$T_\zeta / T_\tau \leqq \frac{4a\zeta_0}{\alpha} \ln \left| \frac{\zeta_0}{\varepsilon} \right|$$

as $z \to z_1^+$, $\zeta_0 \to 0$ and the right-hand side (which is positive) tends to 0. This proves (1).

It then follows that $P(z_k) \to \{\tau = 0\}$ in the sense that all limit points lie in $\{\tau = 0\}$. For (2) it suffices to show that any such limit point is z_1^+. Suppose $P(z_k) \to \tilde{z}$ where $\tilde{z} \neq z_2^-$; then $\tilde{z}$ has coordinates $(-\varepsilon, \tilde{\eta}, 0)$ for some $\tilde{\eta} \neq 0$. But such a point lies in the range of P, so $P(\hat{z}) = \tilde{z}$, for some $\hat{z} \in \{\eta = \varepsilon\} \cap \{\tau = 0\} \cap \{-\varepsilon \leqq \zeta < 0\}$. By continuity of the flow we would then have $z_k \to \hat{z}$, but $\hat{z} \neq z_1^+$ which is a contradiction.

Proof of Lemma 4.2. Recall that $Q_1 : D_1^- \to \{\eta = \varepsilon\}$ and $Q_1(z_1^-) = z_1^+$. We shall compute $\partial Q_1 / \partial w$ in two different directions w. Let $g(z)$ be the vector field of (4.1) at $z = (\zeta, \eta, \rho)$. We claim the lemma follows from proving the following:

(1) $\partial Q_1 / \partial w \times g(z_1^+) = (a_1, a_2, a_3)$ has $a_1 > 0$ and $a_3 > 0$ if $w = (0, 0, 1)$.

(2) $\partial Q_1 / \partial w \times g(z_1^+) = (b_1, b_2, b_3)$ has $b_3 > 0$ if $w = (0, -1, 0)$.

To see that these prove the lemma, consider the images of the curves $\eta = 0$, $\zeta = -\varepsilon$ and $\tau = 0$, $\zeta = -\varepsilon$ inside the neighborhood D_1^-. Q_1 maps these to curves in $\{\eta = \varepsilon\}$ emanating from z_1^+. Further the region $\{\eta \leqq 0\} \cap D_1^-$ is mapped under Q_1 to a region bounded by these two curves, since Q is one-to-one. If D_1^- is small enough, it easily follows from the information about the tangent vectors in (1) and (2) that this region lies in K_a^1 some $a > 0$, see Fig. 13 of $\eta = \varepsilon$. The shaded area is $Q_1(\{\eta \leqq 0\} \cap D_1^-)$. To see this, one has to calculate the cross product with the vector field at this point, but notice the vector field is only nonzero in the second component at z_1^+.

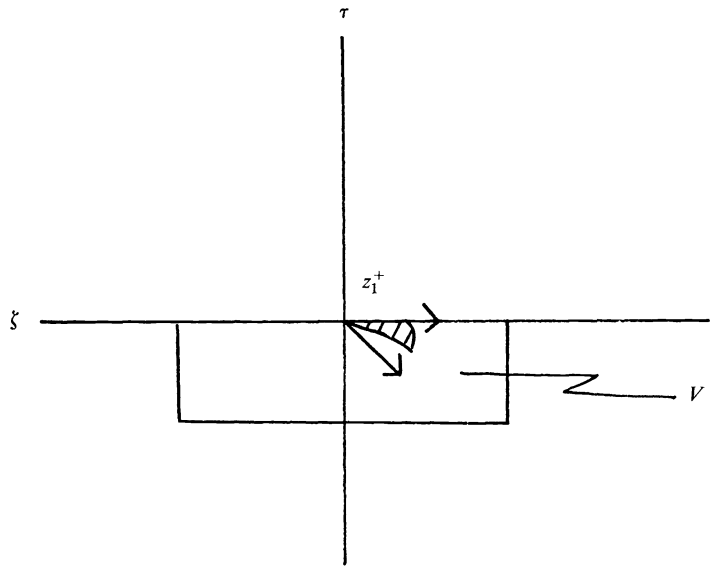


FIG. 13

To prove (1) and (2), calculate:

$$\frac{\partial Q_1}{\partial w} = D_z\Phi(z,\tau(z))w + \frac{\partial \Phi}{\partial t}(z,\tau(z))(\nabla\tau(z)\cdot w);$$

recall that Φ is the flow. Now

$$\frac{\partial Q_1}{\partial t} \times g(z_1^+) = D_z\Phi(z,\tau(z))w \times g(z_1^+) \tag{4.4}$$

since $(\partial\Phi/\partial t)(z_1^-,\tau(z_1^-)) = g(z_1^+)$. So it suffices to compute the right-hand side of (4.4) which is the same as

$$D_z\Phi(z,t)w \times \frac{\partial \Phi}{\partial t}(z,t)\Big|_{(z_1^-,\tau(z_1^-))}. \tag{4.5}$$

We can now see why information about the tangent vectors is obtained through the cross product with $g(z_1^+)$ as in (1) and (2). The problem is that Q_1 is not exactly the flow, but it differs from the flow by something in the direction of the flow and this drops out when crossed with the vector field. (4.5) can be computed using the equation of variations.

Both $D_z\Phi(z,t)w$ and $(\partial\Phi/\partial t)(z,t)$ satisfy the equation of variation of (4.1). It is most convenient to work with this back in (u,v,ρ) variables. It is:

$$\begin{pmatrix} \delta u \\ \delta v \\ \delta\rho \end{pmatrix}' = \begin{pmatrix} 0 & \rho & v \\ -\rho f'(u) & -(n-1)(1-\rho) & (n-1)v - f(u) \\ 0 & 0 & (1-\rho)^2 - 2\rho(1-\rho) \end{pmatrix} \begin{pmatrix} \delta u \\ \delta v \\ \delta\rho \end{pmatrix}. \tag{4.6}$$

Since the orbit through z_1^+ has $\rho = 1$, we can simplify (4.6) to:

$$\begin{aligned} \delta u' &= \delta v + v\delta\rho, \\ \delta v' &= -f'(u)\delta u + \{(n-1)v - f(u)\}\delta\rho, \\ \delta\rho' &= 0. \end{aligned} \tag{4.7}$$

In these co-ordinates, we can write

$$\begin{aligned} D_z\Phi(z,t) &= (\delta u_2(t), \delta v_2(t), c), \\ \frac{\partial \Phi}{\partial t}(z,t) &= (\delta u_1(t), \delta v_1(t), 0) \end{aligned}$$

where $c = -1$ for (1) and $c = 0$ for (2). Note that c is independent of t because of $\delta\rho' = 0$ in (4.7). Now compute

$$D_z\Phi(z,t) \times \frac{\partial \Phi}{\partial t}(z,t) = (-(c\delta v_1), (c\delta u_1), (\delta u_2\delta v_1 - \delta v_2\delta u_1)). \tag{4.8}$$

With $c = -1$, projecting into $\{\tau = 0\}$, (4.8) has co-ordinates $(\delta v_1, -\delta u_1)$. In the $\tau = 0$ plane this is orthogonal to the vector field at z_1^+, in a counterclockwise direction. Since the transformation to (ζ,η) co-ordinates is orientation preserving, (a_1,a_2) lies in a counterclockwise direction from $g(z_1^+)$ and so $a_1 > 0$.

We need to show that for each of the values of w: $(0,0,1)$ and $(0,-1,0)$, the third component is negative, since τ reverses the orientation of ρ. Let $\omega = \delta u_2\delta v_1 - \delta v_2\delta u_1$ and compute, using (4.7),

$$\omega' = -cv^2(n-1).$$

In case (1), $c=-1$ and $\omega'>0$. For (2) $\omega'=0$. It remains to compute the values at $t=0$. This can be done back in (ζ,η,τ) co-ordinates. Since the change of co-ordinates preserves orientation in $\tau=0$, the sign of ω is the same in each co-ordinate system.

For (1), ω is the third component of $(0,0,1)\times g(z_1^-)$, which is zero. For (2), we have $(0,-1,0)\times g(z_1^-)$ and $g(z_1^-)$ has zero second and third components. Moreover its first component is positive. It follows that $\omega<0$ at $t=0$.

In both cases at $t=\tau(z_1^-)$, $\omega<0$ as desired.

Proof of Lemma. 4.1. Using these results, we can prove Lemma 4.1. It suffices to find for given N, $\tilde{x}\in\Gamma^0$ so that $\tilde{x}\cdot[-T,0]\subset W$ and if $\tilde{x}\cdot t=(u(t), v(t), \rho(t))$ then $u(t)$ has at least N zeros in $[-T,0]$. Clearly there exists a smallest $\tilde{T}>T$ so that $\tilde{x}\cdot\tilde{T}\in\{\rho=1-\gamma\}$, set $\tilde{x}\cdot(-\tilde{T})=x$. From Proposition 3.5, $I(C_x)=$number of zeros of $u(t)$, $t\in[-\tilde{T},0]$, which is bigger than N.

To find such an $\tilde{x}$, work in (δ,η,τ) variables. Recall Γ is a curve in W^{cs}_{loc} joining W^{s}_{loc} and some point on the ρ-axis, not $(0,0,1)$. In (δ,η,τ) variables we can assume this includes a piece of $\{\eta=0\}\cap V$ and this piece is a curve converging to the ζ-axis. It follows that we can pick $z\in\Gamma^0$ so that, for some $t_0<0$, $z\cdot t_0\in D_1^-\cap\{\eta\leqq 0\}$; for any neighborhood D_1^- of z_1^-, obviously z depends on D_1^-. Now apply Q_1, by Lemma 4.2 $Q_1(z\cdot t_0)\in K_a^1$ for some $z>0$, and by continuity it lies in a prescribed D_1^+, neighborhood of z_1^+. Therefore there exists t_1, so that $z\cdot t_1\in K_a^1\cap D_1^+$. Now apply P, through Lemma 4.2, $P(z\cdot t_1)\in\{\delta=-\varepsilon\}\cap\{\eta\geqq 0\}\cap D_2^-$. So there exists t_2 such that $z\cdot t_2\in\{\delta=-\varepsilon\}\cap\{\eta\leqq 0\}\cap D_2^-$.

We now have a z so that the orbit $z\cdot[t_2,0]$ rotates around one half-time, through a neighborhood of z_1^- to a neighborhood of z_2^-. It is also clear that $\tau<0$ on this orbit. Notice the important fact that back in (u,v,ρ) co-ordinates $u(t)$ must have a zero between the neighborhood of z_1^+ and that of z_2^-.

Lemma 4.3 can now be applied to $z\cdot t_2$, to obtain t_3 so that $z\cdot t_3\in K_a^2\cap D_2^+$, obviously making D_1^- smaller again if necessary. Then apply Lemma 4.5 to obtain t_4 so that $z\cdot t_4\in\{\delta=\varepsilon\}\cap\{\eta\leqq 0\}\cap D_1^-$, with a possibly different D_1^-. Now a full circuit has been made and $u(t)$ has two zeros. This argument can now be repeated to obtain an arbitrarily large number of zeros.

5. Completion of proof when $\sigma<2/(n-2)$. Recall that we are working with the equation (2.4) in the phase space $\mathbb{R}^2\times[0,1]$. So long as $n>1$, Lemma 4.1 guarantees the existence of a curve $C_x\subset R(\Gamma^0)\subset W^{cs}$ with $I(C_x)$ arbitrarily large. This lies in the set $\{\rho=1-\gamma\}$. The flow operator (3.5) can then be applied to this set. We shall apply it with large negative t so as to pull W^{cs} back to $\rho=0$. We need an estimate to guarantee that Φ is defined and to see how $\Phi_t(C_x)$ can grow in $\mathbb{R}^2$.

It suffices to prove the estimate for the original equation.

LEMMA 5.1. *If $u(r)$ is a solution (with $n>2$) of*

$$u''+\frac{n-1}{r}u'+f(u)=0,\qquad u(r_0)=u_0,\quad u'(r_0)=v_0 \tag{5.1}$$

where $f(u)$ is C^1 and $F(u)=\int_0^u f(s)\,ds$ is bounded below, then there exist C_1, C_2 so that, if $r<r_0$

$$|u(r)|\leqq C_1 r^{2-n} \tag{5.2}$$

and

$$|u'(r)|\leqq C_2 r^{1-n} \tag{5.3}$$

where C_1, C_2 are constants depending on (u_0, v_0), uniformly on compact sets of (u_0, v_0) values for fixed $r_0 \in (0, \infty)$.

Proof. Rewrite (5.1) as

$$(r^{n-1}u')' + r^{n-1}f(u) = 0;$$

multiply by $r^{n-1}u'$,

$$\left\{\frac{1}{2}(r^{n-1}u')^2\right\}' + r^{2n-2}f(u)u' = 0.$$

Now integrate from r_0 to r,

$$\frac{1}{2}(r^{n-1}u')^2 - \frac{1}{2}(r_0^{n-1}v_0)^2 + \int_{r_0}^{r} r^{2n-2}f(u)u' = 0.$$

Integrating by parts, with $F(u) = \int_0^u f(s)\,ds$, gives

$$\frac{1}{2}(r^{n-1}u')^2 = \frac{1}{2}(r_0^{n-1}v_0)^2 - r^{2n-2}F(u)\Big|_{r_0}^{r} + (2n-2)\int_{r_0}^{r} F(u(s))s^{2n-3}\,ds.$$

Since $F(u)$ is bounded below, we can find $k > 0$ so that $F(u) \geqq -k$; then if $r \leqq r_0$

$$\frac{1}{2}(r^{n-1}u')^2 \leqq k_1 + kr^{2n-2} + k(2n-2)\int_r^{r_0} s^{2n-3}\,ds \leqq k_2 \tag{5.4}$$

where k_1, k_2 depend on u_0, v_0 and r_0. From this, (5.3) easily follows and an integration from r to r_0 gives (5.2).

COROLLARY 5.1. *Under the same assumptions as in Lemma* 5.1, *if* $(u(t), v(t), \rho(t))$ *satisfy* (2.4) *with* $(u(0), v(0), \rho(0)) = (u_0, v_0, \rho_0)$, $0 < \rho_0 < 1$ *then there exist constants* C_1, C_2 *such that*

$$|u(t)| \leqq C_1[\rho(t)]^{2-n}, \tag{5.5}$$

$$|v(t)| \leqq C_2[\rho(t)]^{1-n} \tag{5.6}$$

for all $t \in (-\infty, 0]$, *where* C_1 *and* C_2 *depend uniformly on compact sets of* (u_0, v_0) *values with fixed* ρ_0.

Proof. Lemma 5.1 gives the estimate on solutions of (2.1) and therefore on solutions of (2.3) with $\rho = r/(r+1)$. Now reset the parametrisation to get the appropriate estimate on solutions of (2.4).

If $n = 2$, the above still holds but now with $|\ln r|$ replacing r^{2-n}.

LEMMA 5.2. *If* $u(r)$ *is a solution of*

$$u'' + \frac{1}{r}u' + f(u) = 0, \qquad u(r_0) = u_0, \quad u'(r_0) = v_0 \tag{5.7}$$

where $f(u)$ *is* C^1 *and* $F(u) = \int_0^u f(s)\,ds$ *is bounded below, then there exists* C_1, C_2 *so that if* $r \leqq r_0$

$$|u(r)| \leqq C_1|\ln r| \tag{5.8}$$

and

$$|u'(r)| \leqq C_2(1/r) \tag{5.9}$$

where C_1 *and* C_2 *depend uniformly on* (u_0, v_0) *over a compact set if* $r_0 \in (0, \infty)$ *is fixed.*

Proof. All the reasoning in the proof of Lemma 5.1 holds up to (5.4). This immediately gives (5.9) and an integration gives (5.8).

COROLLARY 5.2. *With the assumptions of Lemma* 5.2 *and the set-up of Corollary* 5.1, *the estimates*

$$|u(t)| \leqq C_1 |\ln \rho(t)|, \tag{5.10}$$

$$|v(t)| \leqq C_2 (1/\rho(t)) \tag{5.11}$$

hold under the same conditions as in Corollary 5.1.

It now follows from Corollaries 5.1 and 5.2 and the equation $\rho' = \rho(1-\rho)^2$ that, so long as $n > 1$, the flow operator (3.5) of (2.4) can be applied to the $\{\rho = 1-\gamma\}$ slice for $t \in [T, 0]$, and $T < 0$. Moreover

$$I(\Phi_t(C_x)) \geqq I(C_x) \tag{5.12}$$

from Proposition 3.6.

Because of the growth condition (2) on $f(u)$, $F(u)$ is bounded below and the above corollaries can be applied. If $n > 2$, it follows that $\Phi_t(C_x)$ lies in a set in $\mathbb{R}^2$ of the form

$$\left[-C_1\rho^{2-n}, C_1\rho^{2-n}\right] \times \left[-C_2\rho^{1-n}, C_2\rho^{1-n}\right],$$

and if $n = 2$ in a set of the form

$$\left[-C_1|\ln\rho|, C_1|\ln\rho|\right] \times \left[-C_2(1/\rho), C_2(1/\rho)\right].$$

We must now analyse the manifold at the other end. Pick $u_0 > 0$ so that if $u \geqq u_0$

$$k_1 u^{\sigma+1} \leqq f(u) \leqq k_2 u^{\sigma+1}; \tag{5.13}$$

this can clearly be done from the conditions on $f(u)$. Now choose ρ_0 so that $W_0^{cu} \cap \{\rho = \rho_0\}$ contains a curve L_0 with $(0,0) \in L_0$, $L_0 \subset \{y \geqq 0\}$ and if $(u(t), v(t), \rho(t))$ is the nontrivial solution of (2.4) with $u(t) \to u_0$ as $t \to -\infty$, $\rho(T_0) = \rho_0$ then $u(T_0) \in L_0$. This can be done as W_0^{cu} can be constructed by a finite number of local center-unstable manifolds out to u_0. We can now state the lemma we need to prove.

LEMMA 5.3. *For each* $\rho \in [0, \rho_0]$, *there is a connected curve* $K_\rho \subset W_0^{cu}$ *with* ρ *fixed on it such that*

(1) $K_\rho \subset \{u \geqq 0\}$,

(2) $(0,0) \in K_\rho$,

(3) $K_\rho \cap \{u = c\rho^{-2/\sigma}\} \neq \varnothing$ *for some* $C > 0$ *independent of* ρ.

Proof. Return to the original system again:

$$u'' + \frac{n-1}{r} u' + f(u) = 0, \tag{5.14}$$

and let

$$u(\gamma, 0) = \gamma. \tag{5.15}$$

For each fixed r sufficiently small, either

$$u(\gamma, r) > \frac{\gamma}{2} \tag{5.16}$$

for all $\gamma \in [u_0, \infty)$, or there exists a $\tilde{\gamma}$ such that

$$u(\tilde{\gamma}, r) = \frac{\tilde{\gamma}}{2} \tag{5.17}$$

and (5.16) holds for all $\gamma \in [u_0, \tilde{\gamma})$. If (5.16) is false, then choosing a minimal γ for which (5.17) holds, continuity of $u(\gamma, r)$ guarantees that (5.16) holds on $[u_0, \tilde{\gamma})$, as it does at u_0.

Corresponding to $u(\gamma, r)$ there is a solution of (2.4), $(u(t), v(t), \rho(t))$ with $\lim_{t \to -\infty} u(t) = u_0$. Set $R_0 = \rho_0/(1-\rho_0)$. For fixed $r \in (0, R_0]$ if (5.16) happens for all $\gamma \in [u_0, \infty)$ then the lemma obviously holds at the associated value of ρ.

Suppose then that there exists a γ for which

$$u(\gamma, r) = \gamma. \tag{5.18}$$

The lemma will follow from finding a constant C for which

$$\gamma \geqq C r^{-2/\sigma}. \tag{5.19}$$

Using the usual Green's function, we can rewrite (if $n > 2$)

$$u(\gamma, r) = \gamma - \left(\int_0^r s\left(1 - (s/r)^{n-2}\right) f(u(\gamma, s))\, ds \right) / (n-2). \tag{5.20}$$

If (5.18) holds, then

$$\frac{\gamma}{2} = \left(\int_0^r s\left(1 - (s/r)^{n-2}\right) f(u(\gamma, s))\, ds \right) / (n-2).$$

Assuming that $\gamma \geqq u_0$

$$\frac{\gamma}{2} \leqq \left(\int_0^r s\left(1 - (s/r)^{n-2}\right) ds \right) \left(k_2 \gamma^{\sigma+1} / (n-2) \right),$$

from which it follows that

$$r^2 \gamma^{\sigma} \geqq n / k_2,$$

and the lemma holds. If $n = 2$, (5.20) is replaced by

$$u(\gamma, r) = \gamma - \int_0^r s(\ln - \ln s) f(u(\gamma, s))\, ds;$$

after repeating the above manipulations we arrive at the same estimate,

$$r^2 \gamma^{\sigma} \geqq 2 / k_2.$$

Translating back to (u, v, ρ) space, the lemma is proved.

We now have enough information to prove the theorem in the cases

(1) $\sigma < 2/(n-2)$, $n > 2$,

(2) $n = 2$.

For case (1), to construct the solution with m zeros where m is even, apply the flow to a curve C_x with $I(C_x) > m$. Choose t large negative so that $\rho(t) \leqq \rho_0$, then $\Phi_t(C_x)$ lies in the region where K_ρ is defined. Now $\Phi_t(C_x)$ lies in a strip about the

v-axis of width $C_1[\rho(t)]^{2-n}$ from (5.5) and K_ρ extends out at least to $C\rho^{-2/\sigma}$ (Lemma 5.3). Set $\rho=\rho(t)$, and then the condition that these intersect if t is large enough negative is that $2-n>-2/\sigma$, or $\sigma<2/(n-2)$, see Fig. 14.

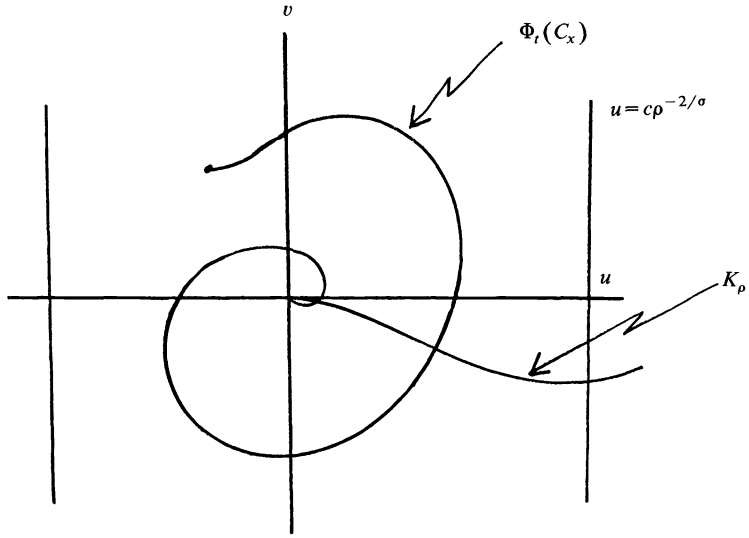


FIG. 14

To check the winding, $I(\Phi_t(C_x))>m$ and so it must intersect K_ρ at a point y so that the piece of $\Phi_t(C_x)$ between y and $(0,0)$ has winding number j, for each j even between 0 and M.

To be more precise, choose t so that

$$C_1[\rho(t)]^{2-n}<C[\rho(t)]^{-2/\sigma}, \tag{5.21}$$

which can be done if $\sigma<2/(n-2)$. Let $\phi(s)$, $s\in[s_0,s_1]$ be a parametrisation of $\Phi_t(C_x)$. Denote by C_s the curve ϕ restricted to $[s_0,s]$. We can find $q=q(j)$ so that $\phi(q)\in\{u=0\}$ and $I(C_q)=j$ for all $0<j\leqq m$. Moreover

$$I(C_s)\leqq j$$

for $s\in[q(j),q(j+1)]$. Let $j<m$ and even, then ϕ restricted to $[q(j),q(j+1)]$ is a curve connecting the negative v-axis to the positive v-axis with the angle making a net increase. Also it stays inside the strip

$$-C_1[\rho(t)]^{2-n}\leqq u\leqq C_1[\rho(t)]^{2-n}.$$

From (5.21), the strip

$$0\leqq u\leqq C_1[\rho(t)]^{2-n}$$

is divided into two components by the curve $K_{\rho(t)}$. In order to increase its angle, therefore, $\phi(s)$ must cross $K_{\rho(t)}$ for some $s\in[q(j),q(j+1)]$. We thus have an intersection of $\Phi_t(C_x)$ with K_ρ at winding number j.

It follows from Proposition 3.5 that the solution starting at this point has exactly j zeros. One checks easily that ρ_0 can be chosen so that it picks up no more zeros on $[0, R_0]$.

For the case of no zeros, we must analyse the tangent vectors of $\Phi_t(C_x)$ and K_ρ at the origin. The tangent vector to $\Phi_t(C_x)$ is always sandwiched between the negative v-axis and the unstable manifold in quadrant four for the $n=1$ case, see §3. However, the tangent vector to K_ρ becomes tangent to the u-axis and so an intersection is forced again if t is large enough negative.

To obtain the solutions with an odd number of zeros, one constructs the curve K_ρ in the left half-plane with the symmetric properties.

For the case $n=2$, observe that the analogue to (5.21) is

$$C_1|\ln \rho(t)| < C[\rho(t)]^{-2/\sigma}$$

which can be satisfied for any σ if t is large enough negative, making ρ small.

6. Completion of proof when $2/(n-2) \leqq \sigma < 4/(n-2)$, $n \neq 2$. The proof of §5 will only work when $\sigma < 2/(n-2)$. It is tempting to try to improve the estimates to prove the full theorem. However, this is not possible; we shall see by the end of this section that the ones given in §5 are, in some sense, the best possible.

The idea of the above proof was to see how far out in the u-direction a slice, in some $\{\rho = \bar{\rho}\}$ plane, of the manifold W_0^{cu} can be extended. In terms of the original problem (1.3), the solutions that satisfy (1.4) end up as the curves on this manifold. If their graphs are drawn u against r, it is well known that these graphs form an envelope, see Fig. 15. In other words they do not fill out the (u,r) plane, at least now while staying positive. Back in (u,v,ρ) space, we can imagine an envelope that marks the edge of the manifold W_0^{cu}. If one can guarantee that some backward iteration of W_{loc}^{cs} gets inside this envelope, existence could be proved, see Fig. 16.

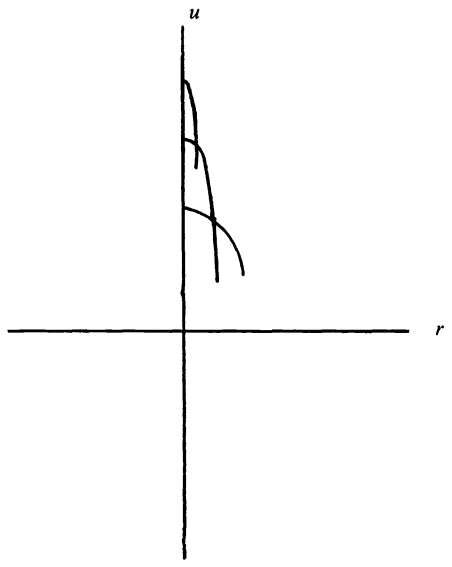


FIG. 15

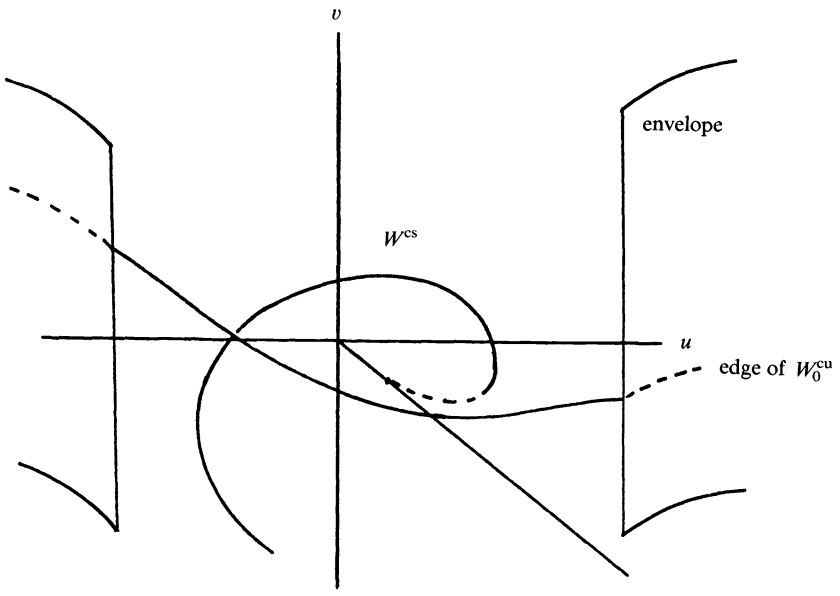


FIG. 16

However, the best general growth estimate one can obtain for these equations is r^{2-n} and the envelope actually grows at the rate $r^{-2/\sigma}$. Hence it looks as if the result of §5 is the best possible. The paradox is resolved by realising that W_0^{cu} does not end at the envelope!

To understand the behavior of W_0^{cu}, we must analyse the limit $r=0$ more carefully. In fact we shall scale so as the envelope stays bounded.

The transformation needed is the Emden–Fowler transformation. This has been used to analyse the pure power and similar cases, see Fowler [7], Chandrasekhar [4], Joseph and Lundgren [10]. Here we shall retrieve the phase planes obtained by these authors as limit systems of a three-dimensional problem.

Consider again the original equation

$$u''+\frac{n-1}{r}u'+f(u)=0.$$

Transform this by

$$y=r^t u, \qquad r=e^s,$$

where $\tau=2/\sigma$. One obtains

$$\ddot{y}+(n-2-2\tau)\dot{y}+\tau(\tau+2-n)y+h(y)y^{\sigma+1}+r^{2+\tau}g(r^{-\tau}y)=0, \qquad \dot{} = \frac{d}{ds},$$

where

$$h(y)=\begin{cases} k_+ & \text{if } y>0, \\ -k_- & \text{if } y\leqq 0, \end{cases}$$

recalling that $f(u)=h(u)|u|^{\sigma+1}+g(u)$, see §1.

We introduce $p=r^{\alpha}$, where α is chosen so that $0<\alpha<\tau$, and write (6.1) as an autonomous system.

$$\begin{aligned}&\dot{y}=z,\\ &\dot{z}=-(n-2-2\tau)z-\tau(\tau+2-n)y-h(y)y^{\sigma+1}-p^{(2+\tau)/\alpha}g(p^{-\tau/\alpha}y),\\ &\dot{p}=\alpha p.\end{aligned}\tag{6.2}$$

This is defined for $p>0$. Now define $q(p,y)$:

$$q(p,y)=\begin{cases}p^{(2+\tau)/\alpha}g(p^{-\tau/\alpha}y) & \text{if } p>0,\\ 0 & \text{otherwise.}\end{cases}\tag{6.3}$$

Consider the system

$$\begin{aligned}&\dot{y}=z,\\ &\dot{z}=-(n-2-2\tau)z-\tau(\tau+2-n)y-h(y)y^{\sigma+1}-q(p,y),\\ &\dot{p}=\alpha p.\end{aligned}\tag{6.4}$$

Note that $p=0$ corresponds to $r=0$ or $s\to-\infty$. We need to check that the right-hand side is C^1. It suffices to check that q is C^1 near $p=0$.

From the assumptions on g given in §1, the following two quantities

$$g(p^{-\tau/\alpha}y)/p^{-\tau\gamma/\alpha}=H_1(p,y)\tag{6.5}$$

and

$$g'(p^{-\tau/\alpha}y)/p^{-\tau(\gamma-1)/\alpha}=H_2(p,y)\tag{6.6}$$

are bounded as $p\to0$ and the bound is uniform over compact sets of y values.

To show that q is continuous, it suffices to show that $q(p,y)\to0$ as $p\to0$ and that this is locally uniform in y. But if $p\neq0$

$$q(p,y)=P^{(2+\tau-\tau\gamma)/\alpha}H_1(p,y).$$

Since $\gamma<\sigma+1$, $2+\tau-\tau\gamma>0$, and so $q(p,y)\to0$ as $p\to0$ uniformly on compact sets of y values. It follows that q is continuous at $p=0$.

To check the derivatives, if $p>0$

$$\frac{\partial q}{\partial y}=p^{(2+\tau)/\alpha-\tau/\alpha}g'(p^{-\tau/\alpha}y)=p^{(2+\tau-\tau-\tau(\gamma-1))/\alpha}H_2(p,y).$$

As above $2+\tau-\tau\gamma>0$ and so $\partial q/\partial y\to0$ as $p\to0$.

To compute $\partial q/\partial p$ at $p=0$, the limit

$$\lim_{p\to0}p^{(2+\tau)/\alpha-1}g(p^{-\tau/\alpha}y)$$

must exist.

$$p^{(2+\tau)/\alpha-1}g(p^{-\tau/\alpha}y)=p^{(2+\tau-\tau\gamma-\alpha)/\alpha}H_1(p,y),$$

and so the above limit exists if

$$2+\tau-\tau\gamma>\alpha\tag{6.7}$$

and is 0. Equation (6.7) is possible for some choice of α so long as

$$2+\tau-\tau\gamma>0$$

which is equivalent to $\gamma<\sigma+1$.

$\partial q/\partial p$ will be continuous if it converges to 0 as $p \to 0$. If $p > 0$

$$\frac{\partial q}{\partial p} = p^{(2+\tau)/\alpha - 1} g\left(p^{-\tau/\alpha} y\right) + p^{2/\alpha - 1} g'\left(p^{-\tau/\alpha} y\right) y.$$

The first term tends to zero as above. The second term is easily seen to converge to zero if (6.7) is satisfied.

This means that the system (6.4) is a C^1 system on $\mathbb{R}^2 \times [0, \infty)$. The plane $p = 0$ is invariant and contains the information about $r = 0$. The next step is to understand the structure of this phase portrait at and near $p = 0$.

The point $(0,0,0)$ is a critical point for (6.4). Linearising at this point, we obtain the matrix

$$\begin{pmatrix} 0 & 1 & 0 \\ -\tau(\tau+2-n) & -(n-2-2\tau) & 0 \\ 0 & 0 & \alpha \end{pmatrix}. \tag{6.8}$$

The eigenvalues of this matrix are τ, $\tau - n + 2$ and α. The associated eigenvectors are $(1, \tau, 0)$, $(1, \tau - n + 2, 0)$ and $(0,0,1)$, respectively. The eigenvalues τ and α are always positive, while

$$\tau - n + 2 > 0$$

if

$$\sigma < 2/(n-2)$$

and $\tau - n + 2 \leqq 0$ if $\sigma \geqq 2/(n-2)$. If we assume $\sigma \geqq 2/(n-2)$, we can form the local unstable manifold $W^{\mathrm{u}}_{\mathrm{loc}}$ to $(0,0,0)$ for (6.4), associated to the two eigenvalues τ and α.

As mentioned at the beginning, if $n = 2$, this method becomes considerably harder to apply but the results of §5 give the theorem.

As usual, set

$$W^{\mathrm{u}} = \bigcup_{t \geqq 0} W^{\mathrm{u}}_{\mathrm{loc}} \cdot t.$$

The reason that this technique works is that W^{u} is mapped back to W^{cu}_0 in (u, v, θ) co-ordinates. The transformation from (y, z, p) to (u, v, ρ), call it G, defined on $\mathbb{R}^2 \times (0, \infty)$ is given by

$$u = p^{-\tau/\alpha} y, \quad v = p^{-(\tau+1)/\alpha}(z - \tau y), \quad \rho = p^{1/\alpha}/\left(1 + p^{1/\alpha}\right).$$

This takes solution curves of (6.4) to those of (2.4).

LEMMA 6.1. $G(W^{\mathrm{u}} \cap \{p \geqq 0\}) \subset W^{\mathrm{cu}}_0$.

Proof. The lemma states that if (y_0, z_0, p_0) is an initial condition for (6.4) with solution $(y(s), z(s), p(s)) \to (0,0,0)$ as $s \to -\infty$, then $G(y_0, z_0, p_0) = (u_0, v_0, \rho_0)$ has a solution $(u(t), v(t), \theta(t))$ to (2.4) with $u(t) \to \gamma$, some $\gamma \in \mathbb{R}$, as $t \to -\infty$. It suffices to show that $u(t)$ is bounded as $t \to -\infty$.

Inside $W^{\mathrm{u}}_{\mathrm{loc}}$, there lies a strong unstable manifold that is one-dimensional; call this $W^{\mathrm{su}}_{\mathrm{loc}}$. It is tangent to the eigenvector $(1, \tau, 0)$ and clearly lies in the invariant plane $p = 0$. Choosing co-ordinates (ζ, η) on $W^{\mathrm{u}}_{\mathrm{loc}}$ so that the p-axis and $W^{\mathrm{su}}_{\mathrm{loc}}$ become axes in that order, it follows easily (see Fenichel [6]) that

$$|\eta(s)| \leqq c e^{\tau s}, \quad \text{some } c > 0$$

as $s \to -\infty$, for any solution on W^{u}_{loc}. Since the eigenvector for τ has a nonzero y component, it follows that

$$|y(s)| \leqq ce^{\tau s}, \quad \text{some } c > 0$$

as $s \to -\infty$. But $u = r^{-\tau} y = e^{-\tau s} y(s)$; so this implies that $|u(s)|$ is bounded as $s \to -\infty$, as desired.

We shall use this lemma to deduce the behavior of W_0^{cu} by understanding W^{u}. The relevant properties about W^{u} will be obtained by the same idea as used in §4, namely to study the limit system and then to view the full problem as a perturbation of this. The next task then is to analyse the system in the $p = 0$ invariant plane.

Restricting (6.4) to $p = 0$,

$$\dot{y} = z, \qquad \dot{z} = (2\tau - n + 2)z + \tau(n - 2 - \tau)y - h(y)y^{\sigma+1}. \tag{6.9}$$

Set $F(y) = \int_0^y (\tau(n-2-\tau)\eta - h(\eta)\eta^{\sigma+1})\,d\eta$ and

$$H(y,z) = \frac{z^2}{2} - F(y); \tag{6.10}$$

then H is the Hamiltonian for (6.9) if the dissipative term is absent, which happens when $2\tau - n + 2 = 0$, or

$$\sigma = 4/(n-2).$$

In general,

$$\dot{H} = (2\tau - n + 2)z^2$$

and $\dot{H} > 0$ if $\sigma < 4/(n-2)$, $\dot{H} < 0$ if $\sigma > 4/(n-2)$. The three-phase portraits are shown in Fig. 17.

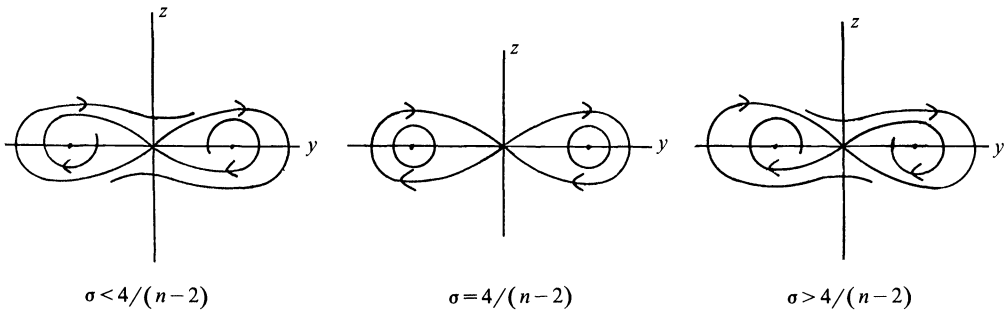


FIG. 17

This is the structure that changes at the critical value of σ. The important point is that the unstable manifold switches from oscillating about the origin to spiralling into one of the right-hand critical points. This is in Joseph and Lundgren [10].

The fact that the unstable manifold spirals out in the subcritical case will perturb to the desired information for W^{u}_{loc} in the full system. The phase portrait depicted for $\sigma < 4/(n-2)$ in Fig. 17 only holds if $\sigma > 2/(n-2)$. At $\sigma = 2/(n-2)$, all those critical points collide at the origin. We need to prove then that this spiralling always occurs if $n \neq 2$.

Let $\tilde{W}^u_{loc}$ be the local unstable manifold at $(0,0)$ for (6.9). Set $\tilde{W}^u = \cup_{s \geqq 0} \tilde{W}^u_{loc} \cdot s$. If $x \in \tilde{W}^u$, there exists an s so that $x \in W^u_{loc} \cdot s$. Let D_x be the curve in $\tilde{W}^u_{loc} \cdot s$ joining x to $(0,0)$. D_x is clearly an admissible curve.

LEMMA 6.2. *If $\sigma < 4/(n-2)$ and $n \neq 2$, then given $m \in \mathbb{Z}^+$, there exists $x \in W^u$ so that $I(D_x) < -m$.*

Proof. $\tilde{W}^u$ is the union of two solutions and $(0,0)$. We must show that each of these solutions oscillates infinitely often about 0 in a clockwise direction. It suffices to show that if $(y(s_0), z(s_0)) \in \{z > 0\} \cap \{H > 0\}$ for some s_0, then there is an $s > s_0$ for which $z(s) = 0$ and at the smallest such $s > s_0$, $y(s) > 0$. This says that any orbit in the upper half plane must leave it at some future time through the positive y-axis. This, together with the symmetric statement about the lower half plane, proves the lemma.

Firstly we prove that if $z(s) \neq 0$ for all $s > s_0$ then $y(s)$ becomes infinite, i.e. given k there exists $s > s_0$ so that $y(s) > k$. Suppose $|y(s)| \leqq k$ for all $s \geqq s_0$; then from (6.9),

$$\frac{d}{ds}\left(e^{-(2\tau - n + 2)s} z\right) \leqq C$$

for some constant $C > 0$. But then $z(s)$ cannot blow up in finite time. Suppose $z(s) \to +\infty$ as $s \to +\infty$; then $y(s) \to +\infty$ as $s \to +\infty$ also. It follows that $z(s)$ stays bounded. But then $\omega(y(s_0), z(s_0))$ is nonempty. Since $\dot{H} > 0$ there are no periodic orbits so, by Poincaré–Bendixson, $\omega(y(s_0), z(s_0))$ must contain a critical point. But since all the critical points lie in $H < 0$ (this is just the point $(0,0)$ if $\sigma < 2/(n-2)$), this contradicts the fact that $H > 0$ on the orbit. Therefore $y(s)$ must become unbounded.

Suppose that for any given k, there exists $\tilde{s}$ so that $y(\tilde{s}) = k^{1/\sigma}$. Set $z = \delta w$ in (6.9) to obtain

$$\dot{y} = \delta w, \qquad \dot{w} = (2\tau - n + 2)w - y\left(|y|^\sigma - \tau(n - 2 - \tau)\right)/\delta. \tag{6.11}$$

Let $\theta = \arctan(w/y)$, computing

$$\dot{\theta} = (2\tau - n + 2)\sin\theta\cos\theta - \cos^2\theta\left(|y|^\sigma - \tau(n - 2 - \tau)\right)/\delta - \delta\sin^2\theta. \tag{6.12}$$

Since $|y|^\sigma > k$

$$\dot{\theta} < (2\tau - n + 2)\sin\theta\cos\theta - (k/\delta)\cos^2\theta - \delta\sin^2\theta + (\tau(n - 2 - \tau)/\delta)\cos^2\theta.$$

Now let $\delta = k^{1/2}$,

$$\dot{\theta} < -\delta^{1/2} + C$$

where C is some constant bounded independently of δ if it is large. But then there exists an s so that $\theta(s) = 0$ by choosing k large enough. This means $z(s) = 0$.

Checking the vector field on $z = 0 \cap \{H > 0\}$ shows that when $z(s) = 0$ for the first time, $y(s) > 0$. This proves the lemma.

In the same spirit as §4, we now show that this winding property can be inherited by $W^u \cap \{p = \gamma\}$ for small enough γ.

Define the Wazewski set

$$W = \{(y, z, p) : p \in [0, \gamma]\}.$$

This is clearly a Wazewski set in forward time if γ is small. Choose γ small enough so that $W^u_{loc} \cap \{p = \gamma\}$ is a one-dimensional curve. Let Γ be a curve $\phi : [\gamma_0, \gamma_1] \to W^u_{loc} \cap \{0 \leqq p \leqq \gamma\}$ such that $\phi(b) \in \{p = 0\}$, $\phi([a, b]) \subset \{y > 0\}$ and $\phi(a) = (0, 0, \gamma)$. Set $\Gamma^0 = \Gamma \setminus \{\phi(b)\}$; then $\Gamma^0 \subset W^0$ since $p' = \alpha p$. If $R : W \to W^-$ is the Wazewski map, set

$\Omega = R(\Gamma^0)$. Obviously Ω is a continuous curve in $\{p=\gamma\}$, so it can be considered as a curve in $\mathbb{R}^2$. With $\psi(s)=R(\phi(s))$, which is defined for $s\in[a,b)$, and $D_s=\psi([0,s))$ we claim the following are true of Ω.

LEMMA 6.3. *Under the assumptions of Lemma* 6.2

(1) $(0,0)\in\Omega$ *and* Ω *is admissible.*

(2) *For all* s, $I(D_s)\leqq 0$.

(3) *Given* $m\in\mathbb{Z}^+$, *there exists an* s *so that* $I(D_s)<-m$.

In order to prove Lemma 6.1, we need a set of propositions that are analogous to Propositions 3.3–3.6. The proofs are essentially the same as in that case, only with appropriate reversal of direction.

Let D_x be a connected curve in $W^u\cap\{p=p_0\}$ with $0\leqq p_0\leqq\sigma$, $\psi:[\gamma_0,\gamma_1]\to\{p=p_0\}$ such that $\psi(b)=x$ and $\psi(a)=(0,0)$.

PROPOSITION 6.1. *Any* D_x *as described above is admissible. Let* $(y(s),\ z(s),\ p(s))$ *be the solution satisfying* $(y(0),\ z(0),\ p(0))=x$.

PROPOSITION 6.2. $-I(D_x)=$ *number of zeros of* $y(s)$ *in* $(-\infty,0]$.

Notice the minus sign, $I(D_x)$ is in fact negative. The flow of (6.4) restricted to $\{p=p_0\}$ any $0\leqq p_0\leqq\gamma$ determines a homotopy

$$\Psi:\mathbb{R}^2\times[a,b]\to\mathbb{R}^2 \tag{6.13}$$

with any $a<b<0$. It is well defined because it is a scaling of the original flow which was well defined.

PROPOSITION 6.3. Ψ *is an admissible homotopy.*

PROPOSITION 6.4. $I(\psi_t(D_x))$ *is a decreasing function of* t.

Proof of Lemma 6.3. (1) is obvious as the tangent vector Ω at $(0,0)$ is a tangent vector to W^u_{loc}. By continuity of the flow if $x=(y(0),\ z(0),\ p(0))$ is close enough to $p=0$ it stays close for an arbitrarily long time. Therefore by Lemma 6.2 it has an arbitrarily large number of zeros; by Proposition 6.2, $-I(D_s)$ can be made arbitrarily large. This proves (3).

To prove (2), suppose this were not true; apply the flow (6.13) ψ_t to D_s. From Proposition 6.4 $I(\psi_t(D_s))>0$ for all $t<0$. But for sufficiently large negative t, $\psi_t(D_s)\subset W^u_{\text{loc}}$ and $I(\psi_t(D_s))=0$. This contradiction proves (2) and therefore the lemma.

The transformation back to (u,v,ρ) co-ordinates, namely G takes Ω to a curve $G(\Omega)$ with the same properties at Ω. In other words $G(\Omega)$ is admissible, $I(G(D_s))\leqq 0$ for all s and there exists s so that $I(G(D_s))$ is an arbitrarily large negative number. This follows because the transformation (6.8) is orientation preserving and maps the z-axis to the v-axis.

Now apply the flow Φ (3.6) of (2.4) to C_x back to $\bar\rho=\gamma^{1/4}/(1+\gamma^{1/\alpha})$. If t is such that $\Phi_t(C_x)\subset\{\rho=\bar\rho\}$, let $E_x=\Phi_t(C_x)$. Intersections between E_x and $G(\Omega)$ are solutions of the problem. We want to find intersections that supply the desired number of zeros.

Let us work in the covering space of $\mathbb{R}^2$ as described in §3 with the covering map given by (3.1). Lift both E_x and $G(\Omega)$ to this space $\tilde X$. Let the lift of $G(\Omega)$ be called Y, choose it so that the end point lies in the interval $[0,1]$ on the x-axis. Y is then a curve, from the above mentioned properties of $G(\Omega)$ (Lemma 6.3), that does not intersect the line $x=1$ but intersects $x=0$ and $x=-m$ for every $m\in\mathbb{Z}^+$.

Now suppose $I(E_x)=k$, and k can be made arbitrarily large. Let E_x be given by $\phi:[s_0,s_1]\to\mathbb{R}^2$. The lift of E_x, say $\tilde E_x$, is determined by the choice of interval in which $\tilde\phi(s_0)$ lives, in other words by the choice of N for which $\tilde\phi(s_0)\in[N,N+1]$.

Now Y divides the covering space $\tilde{X}$ into two components. If $(0,\eta)$ is the end-point of Y on $\{r=0\}$, i.e. the x-axis (r is the radial variable in polar co-ordinates on phase space), then $\{(r,x): r=0,\ x<\eta\}$ lies in one component and each line $x=K$, with $k>0$ lies in the other.

With N, k as above, $\tilde{\phi}(S_1)$ must lie in $k+N\leqq x\leqq k+N+1$. It follows that if $k+N\geqq 1$ and $N\leqq 0$ then this lift of E_x must intersect the curve $G(\Omega)$. Each N supplies a different solution corresponding to such an intersection point, call one determined by a choice of N, J_N. Notice that we can find J_N for $|N|$ arbitrarily large as k can be chosen arbitrarily large. We must prove that the solution J_N has $|N|$ zeros.

LEMMA 6.4. *J_N has $|N|$ zeros.*

Proof. Let δ be the intersection point between $G(\Omega)$ and $\tilde{E}_x$. Let B^+ be the portion of $\tilde{E}_x$ between δ and $\tilde{\phi}(S_0)$; similarly B^- is the portion between δ and $\tilde{\phi}(s_1)$. It is obvious that

$$I(B^+)+I(B^-)=|N|.$$

Let $(u(t),\ v(t),\ \rho(t))$ be the solution of (2.4) satisfying

$$(u(0),\ v(0),\ \rho(0))=\left(p(\zeta),\ p_0^{1/4}\left(1+p_0^{1/4}\right)\right);$$

then $I(B^+)=$ number of zeros of $u(t)$ in $[0,\infty)$ and $I(B^-)=$ number of zeros of $u(t)$ in $(-\infty,0]$. Since N is any element of $\mathbb{Z}^+$, we are done.

Remark. This limiting phase portrait shows that the growth estimates of §5 are the best possible. The eigenvalue other than τ at $(0,0)$ is $\tau-n+2$; coming in at this direction would correspond to growth at the rate r^{2-n}, which is the estimate obtained in §5. The envelope corresponds to a curve that stays bounded in the scaled variables. In fact it is tangent to W^u at the first time it turns around. It therefore grows like $r^{-2/\sigma}$.

Acknowledgments. The first author thanks Prof. Hermann Flaschka for showing him an unpublished manuscript of his and Keener's which was influential on this work. He also thanks Prof. Paul Fife for some helpful conversations on the estimate in §5.

REFERENCES

[1] H. BERESTYCKI AND P. L. LIONS, *Existence of stationary states in nonlinear scalar field equations*, in Bifurcation Phenomena in Mathematical Physics and Related Topics, C. Bardos and D. Bessis, eds., Reidel, Dordrecht, 1980, pp. 269–292.

[2] _____, *Nonlinear scalar field equations*, II. *Existence of infinitely many solutions*, Arch. Rational Mech. Anal., 82 (1983), pp. 347–376.

[3] M. S. BERGER, *On the existence and structure of stationary states for a nonlinear Klein–Gordon equation*, J. Funct. Anal., 9 (1972), pp. 249–261.

[4] S. CHANDRASEKHAR, *An Introduction to the Study of Stellar Structure*, Univ. Chicago Press, Chicago, IL, 1939.

[5] C. CONLEY, *Isolated invariant sets and the Morse index*, CBMS Regional Conference Series in Applied Mathematics 38, American Mathematical Society, Providence, RI, 1978.

[6] N. FENICHEL, *Asymptotic stability with rate conditions*, Indiana Univ. Math. J., 23 (1974), pp. 1109–1137.

[7] R. H. FOWLER, *Further studies of Emden's and similar differential equations*, Quart. J. Math. Oxford Ser. 2 (1931), pp. 259–288.

[8] P. HARTMAN, *Ordinary Differential Equations*, Hartman, Baltimore, MD, 1973.

[9] C. K. R. T. JONES, *Spherically symmetric waves of a reaction diffusion equation*, Technical report no. 2046, Mathematics Research Center, Univ. of Wisconsin, Madison, WI, 1980.

[10] D. D. JOSEPH AND T. S. LUNDGREN, *Quasilinear Dirichlet problems driven by positive sources*, Arch. Rational Mech. Anal., 49 (1972/73), pp. 241–269.

[11] D. W. MCLAUGHLIN, J. V. MOLONEY AND A. C. NEWELL, *Solitary waves as fixed points of infinite-dimensional maps in an optical bistable ring cavity*, Phys. Rev. Lett., 51 (1983), pp. 75–78.

[12] J. R. MUNKRES, *Topology, a First Course*, Prentice-Hall, Inc., Englewood Cliffs, NJ, 1975.

[13] Z. NEHARI, *On a nonlinear differential equation arising in nuclear physics*, Proc. Roy. Irish Acad., 62 (1963), pp. 117–135.

[14] S. I. POHOZAEV, *Eigenfunctions of the equation* $\Delta u+\lambda f(u)=0$, Soviet Math. Dokl., 5 (1965), pp. 1408–1411.

[15] G. H. RYDER, *Boundary value problems for a class of nonlinear differential equations*, Pacific J. Math., 22 (1968), pp. 477–503.

[16] W. A. STRAUSS, *Existence of solitary waves in higher dimensions*, Comm. Math. Phys., 55 (1977), pp. 149–162.

SIAM J. MATH. ANAL.
Vol. 17, No. 4, July 1986

006

A MAXIMUM PRINCIPLE FOR AN ELLIPTIC SYSTEM AND APPLICATIONS TO SEMILINEAR PROBLEMS*

DJAIRO G. DE FIGUEIREDO† AND ENZO MITIDIERI‡

Abstract. The Dirichlet problem in a bounded region for elliptic systems of the form

$$(*) \qquad -\Delta u = f(x,u) - v, \qquad -\Delta v = \delta u - \gamma v$$

is studied. For the question of existence of positive solutions the key ingredient is a maximum principle for a linear elliptic system associated with (*). A priori bounds for the solutions of (*) are proved under various types of growth conditions on f. Variational methods are used to establish the existence of pairs of solutions for (*).

Key words. elliptic systems, maximum principle, a priori estimates, positive solutions, monotone iteration method, mountain pass theorem

AMS(MOS) subject classifications. Primary 35J65, 35J50, 47H15

Introduction. In this paper we propose to discuss the elliptic system

$$(0.1) \qquad -\Delta u = f(x,u) - v, \qquad -\Delta v = \delta u - \gamma v \quad \text{in } \Omega,$$

where Ω is a bounded smooth domain in R^N, $N \geqq 2$, subject to Dirichlet boundary conditions $u = v = 0$ on $\partial\Omega$. The solutions (u,v) of this problem represent steady state solutions of reaction diffusion systems of interest in biology, namely systems of the form

$$(0.2) \qquad u_t = D_1 \Delta u + f(u) - v, \qquad v_t = D_2 \Delta v + \varepsilon(u - \gamma v)$$

where D_1, D_2, ε and γ are positive constants, and one looks for solutions $u(t,x)$, $v(t,x)$ defined in $(0,\infty)\times\Omega$, subject to Dirichlet boundary conditions on $(0,\infty)\times\partial\Omega$. The type of nonlinearities which are of importance in the applications will be described in the Examples I and II below. System (0.2) shows that both species may diffuse. In this sense it is an extension of the well-known FitzHugh–Nagumo system, which serves as a model for nerve conduction, cf. [5] or Hastings [7]. We also mention Koga–Kuramoto [10], where the complete system (0.2) appears and steady state solutions are discussed. There is an extensive bibliography in this subject. We mention three additional papers, which are more closely related to the investigation presented here, namely Rothe–de Mottoni [13], Rothe [14] and Lazer–McKenna [11].

In the applications the constants γ and δ, which appear in system (0.1), are taken to be *positive*. So we shall make this assumption throughout this paper. It follows then that the second equation in (0.1) can be solved for v in terms of u. Let us denote by B its solution operator under Dirichlet boundary conditions. That is, given u, we define

*Received by the editors July 6, 1984, and in revised form February 10, 1985. This work was sponsored by the U. S. Army under contract DAAG29-80-C-0041.

†Departamento de Matemática, Universidade de Brasília, Brasília, Brazil. This research was partially done at the Scuola Internazionale Superiore di Studi Avanzati in Trieste, when the first author held a Guggenheim Fellowship.

‡Istituto di Matematica, Università Degli Studi di Trieste, 34100, Trieste, Italy.

Bu as the solution of the problem $-\Delta v+\gamma v=\delta u$ in Ω, $v=0$ on $\partial\Omega$. Thus our problem becomes the one of finding u such that

$$-\Delta u+Bu=f(x,u) \quad \text{in } \Omega, \qquad u=0 \quad \text{on } \partial\Omega. \tag{0.3}$$

We observe that the left side of (0.3) contains a local (differential) operator $-\Delta$, and a nonlocal (integral) operator B. This fact gives rise to quite interesting questions. It is essential at the outset to understand the operator $-\Delta+B$. In §1 we study its spectral properties and establish a maximum principle for solutions of linear equations like

$$-\Delta u+Bu-\lambda u=g(x) \quad \text{in } \Omega, \qquad u=0 \quad \text{on } \partial\Omega, \tag{0.4}$$

where the real parameter λ is restricted to certain ranges depending on γ, δ and the region Ω. In §2 we establish a priori bounds for solutions of (0.3) under the main assumption that the nonlinearity f at ∞ is below the smallest eigenvalue of the operator $-\Delta+B$; this assumption will be stated precisely as condition (f2) and it characterizes a class of systems which are here called sublinear. The two examples below, which were treated by previous authors [9], [11], [13] and [14], are included in the classes studied in the present paper. Their results are therefore sharpened as far as ranges of the parameters involved and signs of the solutions.

Example I. $f(u)=\lambda u-g(u)$, where λ is a real parameter larger than the first eigenvalue of the operator $-\Delta+B$, and g is a function behaving like u^3, but not necessarily odd; cf. [11], [13], [14].

Example II. $f(u)=u(u-a)(1-u)$, where a is such that $0<a<1/2$. This is the sort of nonlinearity arising in the FitzHugh–Nagumo equations, [5], [9].

The a priori bounds obtained in §2 will be needed in an essential way to perform appropriate truncations of the nonlinearity f, so the problem could be treated by variational methods. This will be done in §5.

In §3 we discuss a class of systems whose model nonlinearity is the one given by Example I. Using the results of §1 we are able to establish the existence of a positive and a negative solution. This result complements a previous one by Lazer and McKenna [11], who proved the existence of two nontrivial solutions by topological degree arguments. Their method, however, does not yield the signs of the solutions obtained. The maximum principle for equations like (0.4) becomes very useful in this respect.

In §4 we sketch a result on the existence of positive solutions for a superlinear elliptic system. Results similar to the ones known for the scalar case hold true in view of the aforementioned maximum principle. The question of the a priori bounds for positive solutions of superlinear elliptic systems may be a hard one. If the growth of the nonlinearity at $+\infty$ is at most like $(N+1)/(N-1)$, for $N\geqq 3$, then the results of Brézis–Turner [2] extend readily. The range $[(N+1)/(N-1), (N+2)/(N-2)]$ poses serious difficulties. The methods used in de Figueiredo–Lions–Nussbaum [3] to treat the scalar case rely on the results of Gidas–Ni–Nirenberg [6], which are not available as yet for the type of systems studied here. We remark that Troy [15] has extended some of the results in [6] to systems. However, Troy's systems do not include the ones we are concerned with. Also in §4 we prove a nonexistence result basing it on our extension to systems of the well-known Pohozaev's identity.

In §5 we consider a class of systems whose model nonlinearity is the one given in Example II. Using the Mountain Pass Theorem of Ambrosetti–Rabinowitz [1], we establish Theorem 5.1 on the existence of two nontrivial solutions for such systems, extending a previous result of Klaasen–Mitidieri [9]. This result shows clearly the relevance of the volume of Ω and of the parameters γ and δ on the existence questions.

It also exhibits the importance of a large positive parameter λ on the existence of two positive solutions for the system

$$-\Delta u=\lambda f(x,u)-v, \qquad -\Delta v=\delta u-\gamma v, \quad \text{in } \Omega$$

subject to Dirichlet boundary conditions, and the nonlinearity f is of the type given by Example II. This relates to the scalar case studied in Rabinowitz [12].

The contents of this paper are as follows:

1. The operator $-\Delta+B$.
2. A priori bounds for solutions of sublinear elliptic systems.
3. Existence of positive solutions.
4. Remarks on a superlinear system.
5. Existence of two nontrivial solutions for a class of sublinear systems.

1. The operator $-\Delta+B$. Consider the linear Dirichlet problem

$$-\Delta v+\gamma v=\delta u \quad \text{in } \Omega, \qquad v=0 \quad \text{in } \partial\Omega, \tag{1.1}$$

where $\Omega\subset R^N$ is a bounded and smooth domain, γ and δ are positive constants. Let us denote by B its solution operator: $v=Bu$. It is well known that

$$B:L^2(\Omega)\to H^2(\Omega)\cap H_0^1(\Omega); \quad B:L^p(\Omega)\to W^{2,p}(\Omega); \quad B:C^\alpha(\bar\Omega)\to C^{2+\alpha}(\bar\Omega).$$

Let us define the operator

$$T\equiv -\Delta+B:L^2(\Omega)\to L^2(\Omega), \quad \text{with } D(T)=H^2(\Omega)\cap H_0^1(\Omega).$$

Clearly T is symmetric, that is, $(Tu_1,u_2)=(u_1,Tu_2)$ for all $u_1,u_2\in D(T)$, where $(\ ,\ )$ denotes the L^2 inner product. Using the L^2 regularity theory, one can prove that T is a closed operator. Let us denote by $0<\lambda_1<\lambda_2\leqq\lambda_3\leqq\cdots$ the eigenvalues of $-\Delta$ under Dirichlet boundary conditions, and by ϕ_k the corresponding eigenfunctions. Then it is easily verified that

$$\hat\lambda_k=\lambda_k+\frac{\delta}{\gamma+\lambda_k}, \qquad k=1,2,\cdots \tag{1.2}$$

are eigenvalues of T. Moreover the same ϕ_k's defined above are their corresponding eigenfunctions. Since (ϕ_k) is a complete orthonormal set in L^2, it is readily shown that the $\hat\lambda_k$'s are the *only* eigenvalues of T. We shall prove in the sequel that in fact the *spectrum* $\sigma(T)$ of T *consists precisely of these eigenvalues*. For each λ in the resolvent set $\rho(T)$ of T, let us denote by $T_\lambda=(T-\lambda I)^{-1}$ its corresponding resolvent operator.

LEMMA 1.1 (a representation formula of the resolvent operator for some values of λ). *Suppose that the real numbers a and b satisfy the following conditions*

$$a>-\lambda_1, \quad \gamma+b>-\lambda_1, \quad b\neq 0, \quad \textit{and} \tag{1.3}$$

$$b\gamma+\delta=ab. \tag{1.4}$$

Then $\lambda=-a-b$ is in the resolvent set $\rho(T)$ and

$$T_\lambda=\left[1-b(\gamma+b-\Delta)^{-1}\right](a-\Delta)^{-1}. \tag{1.5}$$

Proof. With $\lambda=-a-b$, one can write

$$T-\lambda I=(a-\Delta)+b\left[\frac{b\gamma+\delta}{b}-\Delta\right](\gamma-\Delta)^{-1}.$$

Using condition (1.4) above, one obtains

$$T-\lambda I=(a-\Delta)\left[1+b(\gamma-\Delta)^{-1}\right]=(a-\Delta)(\gamma-\Delta)^{-1}(\gamma+b-\Delta).$$

Finally using condition (1.3), it follows that

$$T_\lambda=(\gamma+b-\Delta)^{-1}(\gamma-\Delta)(a-\Delta)^{-1},$$

which readily gives (1.5). □

Remark 1.1. A calculation shows that λ, taken in the ranges indicated below, are representable as $\lambda=-a-b$, with a and b satisfying (1.3) and (1.4):

(i) All $\lambda\leqq-\gamma-2\sqrt{\delta}$. These λ's correspond to $b>0$.

(ii) If $\gamma+\lambda_1>\sqrt{\delta}$, there are some additional values of λ, namely $2\sqrt{\delta}-\gamma\leqq\lambda<\lambda_1+\delta/(\gamma+\lambda_1)$. These λ's correspond to b negative in the range $-\lambda_1-\gamma<b<-\delta/(\gamma+\lambda_1)$.

Remark 1.2 (*monotonicity of the sequence* $\hat{\lambda}_k$). We observe that $\gamma+\lambda_1>\sqrt{\delta}$ implies that $\hat{\lambda}_1<\hat{\lambda}_2\leqq\hat{\lambda}_3\leqq\cdots$. Of course one does not have in general such a monotonicity of the eigenvalues $\hat{\lambda}_k$. Clearly $\gamma+\lambda_1>\sqrt{\delta}$ is not a necessary condition, since it in fact implies the stronger statement that the function $s\mapsto s+\delta/(\gamma+s)$ is monotonically increasing in the whole halfline $[\lambda_1,\infty)$. A necessary and sufficient condition for this monotonicity involves also the second eigenvalue λ_2, namely $\delta<(\gamma+\lambda_1)(\gamma+\lambda_2)$.

COROLLARY 1.2 (compactness of T_λ). *For all* $\lambda\in\rho(T)$, *the resolvent operator* T_λ *is compact.*

Proof. For any $\lambda,\mu\in\rho(T)$ one has the resolvent equation

$$T_\mu-T_\lambda=(\mu-\lambda)T_\mu T_\lambda.$$

So if T_λ is compact for some λ, then it is compact for all λ's in the resolvent set. By the previous lemma T_λ is compact for $\lambda\leqq-\gamma-2\sqrt{\delta}$. □

The following result is an immediate consequence of Lemma 1.1 and Remark 1.1 above.

COROLLARY 1.3 (positiveness of T_λ for some values of λ). *If* $\gamma+\lambda_1>\sqrt{\delta}$, *then* T_λ *is positive for all* $2\sqrt{\delta}-\gamma\leqq\lambda<\hat{\lambda}_1$.

Remark 1.3. The positiveness of T_λ is a *maximum principle* for the equation

$$-\Delta v+Bv-\lambda v=u \quad \text{in } \Omega, \qquad v=0 \quad \text{on } \partial\Omega.$$

It says that if $u\in L^2$ and $u\geqq0$ a.e., then $v\geqq0$ a.e. In fact, it follows from the representation formula (1.5) that a *strong maximum principle* holds. Namely, if $u\in C^0(\Omega)$ and $u\geqq0$ in Ω, then $v>0$ in Ω and the outward normal derivative $(\partial v/\partial\nu)<0$. (Recall that Ω is being assumed to be smooth. So the interior sphere condition is satisfied.)

Remark 1.4. If $\gamma>2\sqrt{\delta}$, then the condition $\gamma+\lambda_1>\sqrt{\delta}$ is automatically satisfied, and Corollary 1.3 says that in this case T_λ is positive for λ in an interval which contains 0. In general one cannot expect that T_0 be positive. Indeed, if $\gamma=\delta=1$, then Corollary 1.3 says that T_λ is positive for $1\leqq\lambda<\hat{\lambda}_1$.

PROPOSITION 1.4. *The spectrum* $\sigma(T)$ *of* T *consists of precisely the eigenvalues* $\hat{\lambda}_k$.

Proof. We have seen above that the point spectrum $P\sigma(T)=\{\hat{\lambda}_k:k=1,2,\cdots\}$. Let $\lambda\notin P\sigma(T)$. Then $T-\lambda I$ is one-to-one. If we show that $T-\lambda I$ is onto, it follows by the Closed Graph Theorem that $\lambda\in\rho(T)$. Thus we claim that the equation $Tu-\lambda u=v$ has a solution u for each given $v\in L^2$. Taking $\mu\in\rho(T)$, we see that this equation is equivalent to $Tu-\mu u=(\lambda-\mu)u+v$, or

$$u=(\lambda-\mu)T_\mu u+T_\mu v. \tag{1.6}$$

By the Fredholm alternative (1.6) is solvable iff the homogeneous equation $u=(\lambda-\mu)T_\mu u$ has only the solution $u=0$. But this is actually the case, since this homogeneous equation is equivalent to $Tu=\lambda u$. Recall that $\lambda\notin P\sigma(T)$. □

Remark 1.5. The above proposition follows also from general results in functional analysis. Namely, since T is a self-adjoint operator, it follows that its residual spectrum $R\sigma(T)$ is empty. Next, since $-\Delta-\lambda$ is Fredholm for every $\lambda\in C$, it follows that $-\Delta+B-\lambda$ is also Fredholm for all $\lambda\in C$. Consequently the continuous spectrum $C\sigma(T)$ is also empty.

Remark 1.6 (*a useful inequality*). Let $\tilde{\lambda}$ denote the smallest of the eigenvalues $\hat{\lambda}_k$. We have seen above that $\tilde{\lambda}=\hat{\lambda}_1$ if $\gamma+\lambda_1>\sqrt{\delta}$. We assert that

$$(Tu,u)\geqq\tilde{\lambda}\|u\|^2_{L^2},\qquad \forall u\in D(T). \tag{1.7}$$

Indeed, since (ϕ_k) is a complete orthonormal set in L^2, we can write $u=\sum\alpha_k\phi_k$ where $\alpha_k=(u,\phi_k)$. So

$$(Tu,u)=\sum\alpha_k(Tu,\phi_k)=\sum\alpha_k(u,T\phi_k)=\sum\alpha_k^2\hat{\lambda}_k,$$

from which the claim follows. A similar argument shows that

$$\int|\nabla u|^2+(Bu,u)\geqq\tilde{\lambda}\|u\|^2_{L^2},\qquad \forall u\in H_0^1. \tag{1.8}$$

Remark 1.7 (*uncoupling of systems and maximum principles*). The usual maximum principle for systems, as well as the maximum principle proved here, seems to be related with the possibility of uncoupling the elliptic system. To make precise our observation, let us look at the linear elliptic system

$$-\Delta u=au+bv+f(x),\qquad -\Delta v=cu+dv+g(x) \tag{1.9}$$

subject to Dirichlet boundary conditions: $u=v=0$ on $\partial\Omega$, where Ω is some bounded domain in R^N, and a, b, c and d are real constants. Suppose that $b\neq 0$ and $c\neq 0$; otherwise the problem trivializes. The uncoupling of system (1.9) is possible if the matrix of the coefficients

$$M=\begin{bmatrix} a & b\\ c & d\end{bmatrix}$$

has two distinct eigenvalues, μ_1 and μ_2. Such a condition is equivalent to

$$(a-d)^2+4bc>0. \tag{1.10}$$

Of course this is the case if b and c both have the same sign. However, to infer the signs of u and v from the signs of the corresponding functions in the uncoupled system, one requires that both b and c be positive. This gives the usual maximum principle for systems. On the other hand, if b and c have opposite signs, the uncoupling is still possible provided a and d "compensate" for the negativeness of bc. Through some calculations one can prove the following result, which essentially gives our maximum principle.

PROPOSITION 1.5. *In addition to* (1.10) *assume that* $bc<0$, $c(a-d)>0$, $\mu_1<\hat{\lambda}_1$ *and* $\mu_2<\hat{\lambda}_1$. *Then if* $f\geqq 0$, $g\geqq 0$ *and* $cf\geqq(a-\mu_1)g$, *it follows that the solutions* u *and* v *of* (1.9) *are positive in* Ω.

2. A priori bounds for solutions of sublinear elliptic systems. Let us consider the elliptic system

$$-\Delta u=f(x,u)-v,\qquad -\Delta v=\delta u-\gamma v\quad \text{in } \Omega, \tag{2.1}$$

where Ω is a bounded smooth domain in R^N, subject to Dirichlet boundary conditions. We always assume that γ and δ are *positive constants*. The nonlinearity f is subject to the following conditions:

(f1) $f:\bar{\Omega}\times R\to R$ is locally Lipschitzian,

(f2) $\limsup_{|s|\to\infty}(f(x,s)/s)<\tilde{\lambda}$ (uniformly in Ω), where $\tilde{\lambda}$ denotes the smallest eigenvalue of the operator $-\Delta+B$ studied in §1.

Condition (f2) characterizes system (2.1) as being *sublinear*.

Examples. 1) $f(u)=\lambda u-h(u)u$, where h is a C^1 function such that $h(0)=0$, $h'(s)s>0$ for all $s\neq 0$ and $\liminf_{s\to\pm\infty}h(s)>\lambda$, (for instance $h(s)=s^2$). This is the case considered in [11] and [14].

2) $f(u)=u(u-a)(1-u)$, where $0<a<1$. This is the type of nonlinearity that appears in the FitzHugh–Nagumo equations; cf. [5], [9].

Remark 2.1. By a solution of (2.1) we mean a classical solution. That is, a pair of functions (u,v) which are in $C^2(\Omega)\cap C^0(\bar{\Omega})$ and which are 0 on $\partial\Omega$. We observe that if, $u,v\in H_0^1(\Omega)\cap C^0(\bar{\Omega})$ satisfy (2.1) in the distribution sense, then by a bootstrap argument it follows that $u,v\in C^{2,\alpha}(\bar{\Omega})$. We remark that in general one cannot drop the hypothesis that u and v are in $C^0(\bar{\Omega})$ in order to be able to bootstrap. However, this would be possible provided one assumes some growth condition on f.

In order to obtain the a priori bound for the solutions of (2.1), we shall assume either one of the conditions below.

(f3) $\lim_{|s|\to\infty}(f(x,s)/|s|^p)=0$, where $1<p<(N+2)/(N-2)$, if $N\geq 3$, and $1<p<\infty$, if $N=2$,

(f4) $\limsup_{|s|\to\infty}(f(x,s)/s)<-\delta/\gamma$,

where the limits are uniform in Ω.

Remark 2.2. In the scalar case (i.e. $-\Delta u=f(x,u)$) condition (f4) corresponds to $f(x,s)<0$ for $s>\beta>0$ and $f(x,s)>0$ for $s<-\beta$, where β is some real number.

PROPOSITION 2.1. *Under hypotheses* (f1), (f2) *and* (f3), *the solutions of* (2.1) *are a priori bounded in* L^∞.

Proof. It follows from (f2) that there exist $0<\mu<\tilde{\lambda}$ and $M>0$ such that

$$f(x,s)\leq\mu s+M,\quad \text{for } 0\leq s<\infty,\qquad f(x,s)\geq\mu s-M\quad \text{for } -\infty<s<0. \tag{2.2}$$

The second equation in (2.1) can be solved for v in terms of u. And in this way system (2.1) is equivalent to the equation

$$-\Delta u+Bu=f(x,u), \tag{2.3}$$

using the notation of §1. So we need only to prove bounds on u. The corresponding bounds on v are obtained immediately from the second equation in (2.1). Multiplying (2.3) by u, integrating by parts and using (1.8), we obtain

$$\tilde{\lambda}\int u^2\leq\int|\nabla u|^2+\int(Bu)u=\int f(x,u)u. \tag{2.4}$$

Next we estimate the last term in (2.4) using (2.2)

$$\int f(x,u)u\leq\mu\int u^2+M\int|u| \tag{2.5}$$

which implies $\int u^2\leq C$. (We shall use the same C to denote different constants.) Using (2.4) and (2.5) again and recalling that B is a bounded linear operator in L^2, we conclude that $\int|\nabla u|^2\leq C$. It follows from (f3) that given $\varepsilon>0$ there exists $C_\varepsilon>0$ such that

$$|f(x,s)|\leq\varepsilon|s|^p+C_\varepsilon.$$

Finally using this inequality and invoking L^p estimates and the Sobolev imbedding theorem, we conclude that there exist a constant $\hat{C}$ such that $\|u\|_{L^\infty} \leqq \hat{C}$. □

Remark 2.3. We emphasize that the dependence of $\hat{C}$ on f is through the constants μ, M and C_ε. So if we change f for $|s| \geqq \hat{C}$ maintaining μ, M and C_ε, the new equation (2.3) with this modified f has the same solutions of the original equation (2.3). This fact will be used in §5.

The following result was proved by Rothe [14] and Lazer and McKenna [11] under less general hypotheses on f. The main idea in the proof below is taken from those papers.

PROPOSITION 2.2. *Under hypotheses* (f1), (f2) *and* (f4), *the solutions of* (2.1) *are a priori bounded in* L^∞.

Proof. (i) We first claim that for $u \in C^0(\bar{\Omega})$, with $u = 0$ on $\partial\Omega$, one has

$$\frac{\delta}{\gamma} \min u \leqq (Bu)(x) \leqq \frac{\delta}{\gamma} \max u, \qquad x \in \Omega. \tag{2.6}$$

Indeed we know that $v = Bu$ satisfies the equation

$$v = \frac{1}{\gamma}\Delta v + \frac{\delta}{\gamma} u. \tag{2.7}$$

Let us prove the first inequality in (2.6). If $v \geqq 0$ that inequality is trivially true. So let us assume that for $x_1 \in \Omega$ we have $v(x_1) = \min v < 0$. Then $\Delta v(x_1) \geqq 0$ and (2.7) implies that $v(x_1) \geqq (\delta/\gamma) u(x_1)$, from which the first inequality in (2.6) follows readily. In a similar way we prove the second inequality in (2.6).

(ii) It follows from (f4) that there exist positive constants k and m such that

$$\frac{f(x,s)}{s} \leqq -k < -\frac{\delta}{\gamma}, \qquad |s| \geqq m. \tag{2.8}$$

We claim that $\|u\|_{L^\infty} \leqq m$ for all solutions u of (2.1). Indeed, suppose by contradiction that $\|u\|_{L^\infty} = M > m$ for some solution u. It follows from (2.6), using the first equation in (2.1) that

$$\frac{\delta}{\gamma} \min u \leqq \Delta u + f(x,u) \leqq \frac{\delta}{\gamma} \max u. \tag{2.9}$$

If there is $x_0 \in \Omega$ such that $u(x_0) = M$, we obtain from (2.9) and (2.8) that

$$-\frac{\delta}{\gamma} M \leqq f(x_0, u(x_0)) \leqq -k u(x_0) = -kM,$$

which is impossible. In a similar way we arrive to a contradiction if $u(x) = -M$ for some $x \in \Omega$. □

Next we discuss the question of bounds for positive solutions of the system (2.1). As remarked before we need only to obtain bounds on u, and then corresponding bounds on v follow readily.

PROPOSITION 2.3. *In addition to* (f1) *assume the following condition*:

(f5) *there exists a constant* $m > 0$ *such that* $f(x,s) = 0$ *for* $s \geqq m$.

Then all nonnegative solutions u of (2.3) *are bounded above by m.*

Proof. Given a solution u of (2.1), define the function ω as $\omega(x) = u(x) - m$ for $u(x) > m$ and $\omega(x) = 0$ for $u(x) \leqq m$. Such a ω belongs to $H_0^1(\Omega)$. So it follows from (2.3) that

$$\int |\nabla \omega|^2 + (Bu, \omega) = \int f(x,u)\omega. \tag{2.10}$$

In view of (f5) and the fact that $(Bu)(x)>0$ for $x\in\Omega$, we conclude from (2.10) that $\int|\nabla\omega|^2=0$, which implies $\omega=0$. □

Remark 2.4. This proposition will be used as follows. Suppose that the function f is such that there is an $m>0$ for which $f(x,m)=0$. Then we consider system (2.1) with f replaced by a new function $\tilde{f}$ defined as f for $s\leqq m$ and as 0 for $s>m$. If for this new system we could find a nonnegative solution u, then by the proposition above such a u would be indeed a solution of the original system.

Remark 2.5. Now if $f(x,s)=f(s)$ satisfies (f2), $f(0)\geqq 0$ and $f(s)>0$ for $0<s<\varepsilon$, then either there is an $m>0$ such that $f(m)=0$ or f satisfies (f3). In the first case we treat the problem as in the previous remark. In the second case we proceed as in Proposition 2.1 and obtain an a priori bound on positive solutions.

Remark 2.6. Similar statements can be made for nonpositive solutions u.

Remark 2.7. *A sufficient condition for all (eventual) nontrivial solutions of* (2.1) *to be positive.* Assume that $\gamma+\lambda_1>\sqrt{\delta}$ and that $f(x,u)\geqq\alpha u$ for all u, where $-\gamma+2\sqrt{\delta}\leqq\alpha<\hat{\lambda}_1$. Then *the nontrivial solutions* u of (2.1) *are positive in* Ω. From (2.3) we obtain $-\Delta u+Bu\geqq\alpha u$, and the result follows readily by Corollary 1.3.

Remark 2.8. The previous condition applied to Example 2 gives interesting conclusions. Indeed, we can in this case compute explicitly the value of m in (2.8). Then truncate f outside $|s|\geqq m$ in such a way that the new f has derivative equals to $-a$ for $|s|>m$. By Proposition 2.2 the solutions of (2.1) with this new f are the same as the solutions of the original equation. Moreover, from the way the truncation is done, it follows (by a straightforward calculation) that now $f(u)\geqq -au$, (where we are denoting also by f the truncated function) provided $\delta/\gamma<a$. So the previous sufficient condition applies. Summarizing, *the solutions of* (2.1), *in the case of Example* 2, *are positive if*

$$\frac{\delta}{\gamma}<sa\leqq\gamma-2\sqrt{\delta}\,. \tag{2.11}$$

Observe that, if (2.11) is assumed, then the condition $\gamma+\lambda_1>\sqrt{\delta}$ is automatically satisfied, cf. Remark 1.4. We remark that no solution of (2.3) in this example can be nonpositive (i.e. $u\leqq 0$ in Ω). In fact the solutions in general change sign.

3. Existence of positive solutions. We consider again system (2.1) of the previous section or its equivalent expression in the form of equation (2.3). In this section we examine the question of existence of a positive solution under an additional condition on the nonlinearity f at 0. In order to simplify the presentation in the sequel, we suppose that f does not depend on x. The case when f depends also on x can also be treated by the method used here; under appropriate conditions on f similar results may be obtained. So we assume the next condition.

(f6) $\liminf_{s\to 0}(f(s)/s)>\hat{\lambda}_1$.

Examples. Condition (f6) is satisfied, for instance, if (i) $f(0)>0$, or (ii) $f(s)$ is C^1 and $f'(0)>\hat{\lambda}_1$. A special case of (ii) was considered in [11].

THEOREM 3.1. *Assume that* $\gamma+\lambda_1>\sqrt{\delta}$. *In addition to conditions* (f1), (f2) *and* (f6), *suppose that* f *is* C^1 *for* $s\geqq 0$ *and*

$$\inf\{f'(s):0\leqq s<\beta\}\geqq-\gamma+2\sqrt{\delta}\,, \tag{3.1}$$

where $\beta\leqq+\infty$ *is the first positive zero of* $f(s)$. *Then equation* (2.3) *has a positive solution* u, *or equivalently, system* (2.1) *has a pair* (u,v) *of positive solutions.*

Remark 3.1. The hypothesis $\gamma+\lambda_1>\sqrt{\delta}$ in Theorem 3.1 implies that $\tilde{\lambda}=\hat{\lambda}_1$. Recall also that under this hypothesis $-\gamma+2\sqrt{\delta}<\hat{\lambda}_1$, and so we can make use of Corollary 1.3. The condition on the differentiability of f can be relaxed and in consequence (3.1) has to be replaced by an appropriate one-sided Lipschitz condition.

Proof of Theorem 3.1. (i) It follows from (f6) that there exist $\nu > \hat{\lambda}_1$ and $s_0 > 0$ such that $f(s) \geqq \nu s$ for $0 \leqq s \leqq s_0$. Thus $\varepsilon\phi_1$ is a subsolution of (2.3) for all ε such that $0 < \varepsilon < \varepsilon_0 = s_0 / \max \phi_1$.

(ii) If $\beta < \infty$ then $\omega(x) \equiv \beta$ in Ω is a supersolution of (2.3). If $\beta = +\infty$ we construct a supersolution ω for (2.3) as follows. It follows from (f2) that there exist $-\gamma + 2\sqrt{\delta} < \mu < \hat{\lambda}_1$ and $C > 0$ such that $f(s) \leqq \mu s + C$. We then take ω as the solution of $-\Delta\omega + B\omega = \mu\omega + C$ in Ω, $\omega = 0$ on $\partial\Omega$. In view of Corollary 1.3, $\omega > 0$ in Ω and $\varepsilon > 0$ can be chosen in such a way that $\varepsilon\phi_1 < \omega$ in Ω.

(iii) So (2.3) possesses an ordered pair of a sub- and a supersolution. Now in order to apply the method of monotone iteration, it is still required that (a) $T_\lambda = (-\Delta + B - \lambda I)^{-1}$ be a positive operator for some real number λ, and (b) the function $s \to f(s) - \lambda s$, for the same λ, be nondecreasing in the interval $[0, \max \omega]$. These two requirements are accomplished if one chooses $\lambda = -\gamma + 2\sqrt{\delta}$. Indeed, (a) then follows by Corollary 1.3 and (b) follows from (3.1). Therefore the method of monotone iteration can be applied and one obtains a solution of (2.3) in the interval $[\varepsilon\phi_1, \omega]$.

Remark 3.2. It should be remarked that besides (f2) no growth condition is required on f.

Remark 3.3. A statement similar to Theorem 3.1 holds true for the existence of negative solutions of (2.1). In this case, condition (3.1) is replaced by

$$(3.1)^- \qquad \inf\{f'(s) : \beta' < s \leqq 0\} \geqq -\gamma + 2\sqrt{\delta}$$

where $-\infty \leqq \beta' < 0$ is the first negative zero of $f(s)$. In order to prove such a result, we can reduce it to the situation of Theorem 3.1 by the substitution $z = -u$.

Example. $f(u) = au - u^3$ with $a > 0$. In this case $\beta = \sqrt{a}$, and $\min\{f'(s) : 0 \leqq u \leqq \beta\} = -2a$. So conditions (f6) and (3.1) are satisfied if $\hat{\lambda}_1 < a \leqq \gamma/2 - \sqrt{\delta}$. We then see that in this example there are values of a for which (2.3) has a positive solution provided

$$(3.2) \qquad \hat{\lambda}_1 < \gamma/2 - \sqrt{\delta}\,.$$

Clearly this is the case for instance if γ is large. This is also the case if $\gamma > (\sqrt{3} + 1)\sqrt{\delta}$ and Ω is a sufficiently large ball. Indeed, for large balls λ_1 is essentially zero and this last inequality implies readily condition (3.2). Clearly in this example there is also a negative solution, namely $-u$, where u is the positive solution.

Comparison with the results of Lazer–McKenna. In [11] the following system is studied

$$(3.3) \qquad -k\Delta u = \lambda u - h(u)u - v, \qquad -\Delta v + v = u \quad \text{in } \Omega$$

subject to Dirichlet boundary conditions. Under certain conditions on k, λ and h it is proved that

$$(3.4) \qquad -k\Delta u + (1 - \Delta)^{-1} u = \lambda u - h(u)u,$$

which is an equivalent form of (3.3), has exactly three solutions. In [11] a topological degree argument is used, which does not give the sign of the two nontrivial solutions. Under essentially the same hypotheses, our Theorem 3.1 says that one of these solutions is positive and the other is negative. Our precise result is the following. We state only the one corresponding to the existence of a positive solution. A similar one can be drawn for the existence of a negative solution.

COROLLARY 3.2. *Under the assumptions below, equation* (3.4) *has a positive solution*:

$$1+\lambda_1 > 1/\sqrt{k}\,, \tag{3.5}$$

$$h\in C^1(R,R), \quad h(0)=0 \quad h'(s)s>0 \quad \forall s\neq 0, \tag{3.6}$$

$$k\lambda_1+\frac{1}{1+\lambda_1}<\lambda, \tag{3.7}$$

$$\sup\{h'(s)s+h(s): 0\leqq s\leqq\beta\}\leqq\lambda+k-2\sqrt{k}\,, \tag{3.8}$$

where β *is the only positive solution of* $h(s)=\lambda$. (*Observe that* β *could be* $+\infty$.)

Remark 3.4. If $h'(s)$ is nondecreasing, then $\beta<\infty$ and (3.8) simplify to $\beta h'(\beta)\leqq k-2\sqrt{k}$. So positive solutions of (3.4) exist if the *diffusion rate k is large.*

4. Remarks on a superlinear system. Consider the elliptic system

$$-\Delta u=f(u)-v, \qquad -\Delta v=\delta u-\gamma v \quad \text{in } \Omega, \tag{4.1}$$

subject to Dirichlet boundary conditions, with $\gamma,\delta>0$ and $-\gamma+2\sqrt{\delta}<0$. Assume the following conditions on the nonlinearity f:

(f1)′ $f: R^+\to R^+$ locally Lipschitzian,

(f7) $\liminf_{s\to+\infty}(f(s)/s)>\hat{\lambda}_1$,

(f8) $\limsup_{s\to 0}(f(s)/s)<\hat{\lambda}_1$,

(f9) $\lim_{s\to+\infty}(f(s)/s^\sigma)=0$ where $1<\sigma\leqq(N+1)/(N-1)$, if $N\geqq 3$ and $1<\sigma<\infty$, if $N=2$.

As seen in the previous section, (4.1) is equivalent to

$$-\Delta u+Bu=f(u). \tag{4.2}$$

Under the hypotheses above we may proceed as in the scalar case (cf. Brézis–Turner [2]) and we prove that (4.2) has a positive solution. Condition (f9) is used to get a priori bounds for the positive solutions of (4.2). We do not know how to proceed in order to obtain such bounds in the case when $(N+1)/(N-1)<\sigma<(N+2)/(N-2)$ and $N\geqq 3$. The results of [3] for the scalar case are not immediately extended to this case. For that purpose, the first step would be to see how the results of Gidas–Ni–Nirenberg [6] look (if at all!) in this case. We remark that the extension obtained by Troy [15] does not cover the type of systems studied in this paper.

Remark 4.1. The condition $-\gamma+2\sqrt{\delta}<0$ is used in order to guarantee that the operator $T_0\equiv(-\Delta+B)^{-1}$ is positive. If this condition is not satisfied, but one has $\gamma+\lambda_1>\sqrt{\delta}$, everything still works provided $\hat{\lambda}_1$ in the right sides of assumptions (f7) and (f8) is replaced by $\hat{\lambda}_1-\gamma+2\sqrt{\delta}$.

Nonexistence of positive solutions in the case when $f(u)=u^p$, *for* $p\geqq(N+2)/(N-2)$ *and* $N\geqq 3$. As in the scalar case this is proved using an identity of the Pohozaev type. The function $f(u)=u^p$ for $u\geqq 0$ is extended as $f(u)=0$ for $u<0$. Then it follows from Remark 2.7 that all eventual solutions u and v of (4.1) are positive in Ω, provided we assume that $-\gamma+2\sqrt{\delta}<0$. Consequently the nonexistence of nontrivial solutions for system (4.1) (in *star-shaped* domains Ω) with such an f follows readily from the two lemmas below.

LEMMA 4.1. *Let u and v be solutions of* (4.1). *Then the following identity holds*

$$2N\int F(u)-(N-2)\int uf(u)-2\int uv-\frac{2}{\delta}\int|\nabla v|^2=\oint(x\cdot\nu)\left[|\nabla u|^2-\frac{1}{\delta}|\nabla v|^2\right] \tag{4.3}$$

where $F(s)=\int_0^s f$ *and* $\int$ *denotes* (*volume*) *integral over* Ω *and* $\oint$ (*surface*) *integral over* $\partial\Omega$. *Here* ν *denotes the outward unit normal.*

LEMMA 4.2. *Let u and v be solutions of* (4.1). *Assume that* $-\gamma+2\sqrt{\delta}<0$. *Then* $u-(1/\sqrt{\delta})v$ *is positive in* Ω *and*

$$\frac{\partial u}{\partial \nu}<\frac{1}{\sqrt{\delta}}\frac{\partial v}{\partial \nu}<0 \quad \text{on } \partial\Omega.$$

To conclude this section, we prove the two lemmas above.

Proof of Lemma 4.1. First we use the general form of Pohozaev's identity for solutions of the $-\Delta u=g(x,u)$ in Ω and $u=0$ on $\partial\Omega$; see [3]. This identity will be applied separately to the first and second equations in (4.1). Observe that for the first equation, $g(x,s)=f(s)-v(x)$, and for the second equation, $g(x,s)=\delta u(x)-\gamma s$. Then we obtain the following two identities

$$2N\int[F(u)-uv]-2\int(x\cdot\nabla v)u-(N-2)\int[f(u)-v]u=\oint(x\cdot\nu)|\nabla u|^2, \tag{4.4}$$

$$2N\int\left[\delta uv-\frac{1}{2}\gamma v^2\right]+2\delta\int(x\cdot\nabla u)v-(N-2)\int[\delta u-\gamma v]v=\oint(x\cdot\nu)|\nabla v|^2. \tag{4.5}$$

(If one prefers to ignore [3], identities (4.4) and (4.5) may be obtained in the standard way Pohozaev's identities are proved. Use the multiplier $x\cdot\nabla u$ in the first equation of (4.1) and $x\cdot\nabla x$ in the second.) It follows from the divergence theorem that

$$\int(x\cdot\nabla v)u+\int(x\cdot\nabla u)v=-N\int uv. \tag{4.6}$$

Next dividing (4.5) through by δ, subtracting the result from (4.4) and using (4.6), we obtain

$$2N\int F(u)-(N-2)\int uf(u)-4\int uv+\frac{2\gamma}{\delta}\int v^2=\oint(x\cdot\nu)\left[|\nabla u|^2-\frac{1}{\delta}|\nabla v|^2\right]. \tag{4.7}$$

Now it follows from the second equation in (4.1) that

$$\int|\nabla v|^2=\delta\int uv-\gamma\int v^2. \tag{4.8}$$

Taking (4.8) into (4.7), we obtain the identity (4.3). □

Proof of Lemma 4.2. It follows from $-\gamma+2\sqrt{\delta}<0$ that there exists a real number k such that $\sqrt{\delta}<k<\gamma-\sqrt{\delta}$. Using (4.1), it is easy to check that

$$(-\Delta+k)\left(u-\frac{1}{\sqrt{\delta}}v\right)\geqq 0 \quad \text{in } \Omega$$

from which the assertion of the lemma follows. Observe that we know that all (eventual) solutions of (4.1) would be positive. □

5. Existence of two nontrivial solutions for a class of sublinear systems. Let us once more consider system (2.1) under conditions (f1), (f2), (f3) or (f4). As in previous sections we discuss, instead of system (2.1), its equivalent form given by equation (2.3). In this section we propose to treat the question of existence of solutions of (2.3) by a variational argument. So we look for the critical points of the functional

$$\Phi(u)=\frac{1}{2}\int|\nabla u|^2+\frac{1}{2}(Bu,u)-\int F(x,u) \tag{5.1}$$

where $F(x,s)=\int_0^s f(x,\xi)\,d\xi$. Although this functional is well defined in H_0^1 if we assume (f3), this is not the case if (f4) is assumed instead. Observe that both (f2) and (f4) restrict f only in one direction. So some truncation has to be done. The existence of a priori bounds on the solutions of (2.3) in either case ((f3) or (f4) assumed), as proved in §2, allows us to truncate the nonlinearity f in such a way that the functional Φ is *well defined* in H_0^1 and it is *bounded from below*. Indeed, if case (f3) is assumed, we choose an appropriate $\tilde{C}>\hat{C}$ and do this truncation for $|s|\geqq\tilde{C}$ (see Proposition 2.1 and Remark 2.3) preserving μ, M and C_ε and in such a way that $\lim_{|s|\to\infty}(f(x,s)/s)=l$ where $0<l<\tilde{\lambda}$. In case we assume (f4) the truncation is done for $|s|\geqq m$ (see Proposition 2.2), and in such a way that $\lim_{|s|\to\infty}(f(x,s)/s)=-k$, where the constant k is given in (2.8). The truncation so done has the very essential feature that the new equation (2.3) with this truncated function has the same solutions as the solutions of the original equation (2.3).

One sees immediately that $\Phi: H_0^1(\Omega)\to R$ is C^1 and

$$(\Phi'(u),\omega)_{H^1}=\int\nabla u\cdot\nabla\omega+\int(Bu)\omega-\int f(x,u)\omega. \tag{5.2}$$

So the critical points of Φ are the H_0^1 solutions of (2.3). By a bootstrap argument it follows that these solutions are in fact in $C^{2,\alpha}(\bar{\Omega})$.

LEMMA 5.1. *The functional Φ defined above satisfies the Palais–Smale condition.*

Proof. (i) In view of Poincaré's inequality we may consider H_0^1 endowed with the inner product $(u,\omega)_{H^1}=\int\nabla u\cdot\nabla\omega$. It is well known that the nonlinear operator $\tilde{f}: H_0^1\to H_0^1$ defined by $(\tilde{f}(u),\omega)_{H^1}=\int f(x,u)\omega$, $\forall\,\omega\in H_0^1$, is compact. (Recall that f has linear growth in view of the truncation.) On the other hand the (linear) operator $\tilde{B}: H_0^1\to H_0^1$ defined by $(\tilde{B}u,\omega)_{H^1}=\int(Bu)\omega$ is also compact. This follows readily from the compact imbedding of H_0^1 in L^2. Consequently $\Phi'=I+\tilde{B}-\tilde{f}$, that is, Φ' is of the form identity + compact operator. Thus to prove the Palais–Smale condition, it is enough to show that any sequence $(u_n\in H_0^1)$ such that $|\Phi(u_n)|\leqq C$ and $\Phi'(u_n)\to 0$ in H_0^1 possesses a subsequence (denoted again by u_n) such that $\|u_n\|_{H^1}\leqq C$.

(ii) It follows from $\Phi'(u_n)\to 0$ that given $\varepsilon_n\downarrow 0$ there exists a subsequence of (u_n) (denoted again by u_n) such that

$$\left|\int\nabla u_n\cdot\nabla\omega+\int(Bu_n)\omega-\int f(x,u_n)\omega\right|\leqq\varepsilon_n\|\omega\|_{H^1}. \tag{5.3}$$

Now using (5.3) with $\omega=u_n$ and estimating with the help of (1.8), we get

$$\tilde{\lambda}\int|u_n|^2\leqq\int f(x,u_n)u_n+\varepsilon_n\|u_n\|_{H^1}. \tag{5.4}$$

From the properties of the truncated f we obtain from (5.4)

$$\int u_n^2\leqq C+C\varepsilon_n\|u_n\|_{H^1}. \tag{5.5}$$

Next from $|\Phi(u_n)|\leqq C$ we infer that

$$\int|\nabla u_n|^2\leqq\int(Bu_n)u_n+2\int|F(x,u_n)|+C \tag{5.6}$$

and finally using the properties of the truncated f, we obtain from (5.6) and (5.5) that $\int|\nabla u_n|^2\leqq C+C\varepsilon_n\|u_n\|_{H^1}$ which proves that $\|u_n\|_{H^1}\leqq C$. □

Remark 5.1. It follows immediately from the previous remarks that system (2.1) has *at least one solution* under hypotheses (f1), (f2), and (f3) or (f4). Indeed, since Φ is C^1 functional, bounded below and satisfying the Palais–Smale condition, it follows that it has a global minimum u_1, $\Phi(u_1)=\inf\{\Phi(u): u\in H_0^1\}$. One cannot expect in general the existence of more solutions. Indeed if $f(u)=\lambda u$ with $\lambda<\tilde{\lambda}$, equation (2.3) in this case has only the trivial solution! So some additional assumption is necessary.

Now we treat a problem which is superlinear at 0, in the sense that the condition below holds:

(f10) f is differentiable at 0, $f(x,0)=0$, and $f'(x,0)<\tilde{\lambda}$.

Example 2 in §2 satisfies condition (f10).

THEOREM 5.2. *Assume conditions* (f1), (f2), (f3) *or* (f4), *and* (f10). *In addition suppose that there exists* $\xi>0$ *such that*

$$F(\xi)\geqq F(s) \quad \forall\, 0\leqq s\leqq \xi, \tag{5.7}$$

$$\frac{2F(\xi)}{\xi^2} > \min\left\{ \frac{1}{R^2}\frac{(1+t)^2}{t^2}\frac{(1+t)^N-1}{2-(1+t)^N}+\frac{\delta}{\gamma}\frac{(1+t)^N}{2-(1+t)^N} : 0<t<2^{1/N}-1 \right\} \tag{5.8}$$

where R *denotes the radius of the largest ball contained in* Ω. *Then equation* (2.3) *has at least two nontrivial solutions.*

Remark 5.2. Condition (5.8) is the analogue of a condition introduced by one of the authors (D.G.F.) in [4] for the scalar case. We remark that if there is a $\xi>0$ such that $F(\xi)>0$ then the condition is satisfied (for example if Ω is a large ball and δ is very small). The special case of Example 2 was studied by Klaasen and Mitidieri [9]. Condition (5.8) follows readily from their conditions: (i) Ω to be a large ball, and (ii) $\gamma/\delta>9/(2a^2-5a+2)$.

Proof. It suffices to prove that there exists $\tilde{u}\in H_0^1$ such that $\Phi(\tilde{u})<0$. Once this is done, we see that the global minimum u_1 of Φ is a nontrivial solution since $\Phi(u_1)=\inf\Phi<0$. The second solution is obtained immediately by an application of the Mountain Pass Theorem of Ambrosetti–Rabinowitz [1], since 0 is a strict local minimum in view of assumption (f10). In order to see that there are points in H_0^1 where the functional Φ is negative we consider the functions u_t below. We may assume that the ball centered at 0 with radius R is contained in Ω, where R is the radius of the largest ball contained in Ω. Defining

$$u_t(x)=\begin{cases} \xi & \text{if } \|x\|\leqq R/(1+t), \\ \xi\left[1-\dfrac{1+t}{tR}\left(\|x\|-\dfrac{R}{1+t}\right)\right] & \text{if } \dfrac{R}{1+t}\leqq\|x\|\leqq R, \\ 0 & \text{if } x\in\Omega\setminus B_R(0), \end{cases}$$

the result follows by a calculation from conditions (5.7) and (5.8). □

REFERENCES

[1] A. AMBROSETTI AND P. H. RABINOWITZ, *Dual variational methods in critical point theory and applications*, J. Funct. Anal., 14 (1973), pp. 349–381.

[2] H. BRÉZIS AND R. L. TURNER, *On a class of superlinear elliptic problems*, Comm. Partial Differential Equations, 2 (1977), pp. 601–614.

[3] D. G. DE FIGUEIREDO, P.-L. LIONS AND R. NUSSBAUM, *A priori estimates and existence of positive solutions of semilinear elliptic equations*, J. Math. Pures Appl., 61 (1982), pp. 41–63.

[4] D. G. DE FIGUEIREDO, *Positive solutions for some classes of semilinear elliptic problems*, to appear.

[5] R. FITZHUGH, *Impulses and physiological states in theoretical models of nerve membrane*, Biophysic. J., 1 (1961), pp. 445–466.

[6] B. GIDAS, W. M. NI AND L. NIRENBERG, *Symmetry and related properties via the maximum principle*, Comm. Math. Phys., 68 (1979), pp. 209–243.

[7] S. P. HASTINGS, *Some mathematical problems from neurobiology*, Amer. Math. Monthly, 82 (1975), pp. 881–895.

[8] G. KLAASEN AND W. TROY, *Standing wave solutions of a system of reaction-diffusion equations derived from FitzHugh–Nagumo equations*, SIAM J. Appl. Math., 44 (1984), pp. 96–110.

[9] G. KLAASEN AND E. MITIDIERI, *Standing wave solutions for a system derived from the FitzHugh–Nagumo equations for nerve conduction*, to appear.

[10] S. KOGA AND Y. KURAMOTO, *Localized patterns in reaction-diffusion systems*, Progr. Theoret. Phys., 63 (1980), pp. 106–121.

[11] A. C. LAZER AND P. J. MCKENNA, *On steady state solutions of a system of reaction-diffusion equations from biology*, Nonlinear Anal., Theory, Meth. Appl., 6 (1982), pp. 523–530.

[12] P. H. RABINOWITZ, *Pairs of positive solutions of nonlinear elliptic partial differential equations*, Indiana Univ. Math. J., 23 (1973), pp. 173–186.

[13] F. ROTHE AND P. DE MOTTONI, *A simple system of reaction-diffusion equations describing morphogenesis: Asymptotic behavior*, Ann. Mat. Pura Appl., 122 (1979), pp. 141–157.

[14] F. ROTHE, *Global existence of branches of stationary solutions for a system of reaction diffusion equations from biology*, Nonlinear Anal., Theory, Meth. Appl., 5 (1981), pp. 487–498.

[15] W. C. TROY, *Symmetry properties in systems of semilinear elliptic equations*, J. Differential Equations, 42 (1981), pp. 400–413.

SIAM J. MATH. ANAL.
Vol. 17, No. 4, July 1986

007

ON THE EIGENVALUE PROBLEM FOR COUPLED ELLIPTIC SYSTEMS*

ROBERT STEPHEN CANTRELL† AND KLAUS SCHMITT‡

Abstract. We consider the eigenvalue problem

$$\begin{aligned} L_k u_k &= \lambda \sum_{i=1}^{r} m_{ki} u_i \quad \text{in } \Omega, \\ u_k &= 0 \qquad\qquad\quad \text{on } \partial\Omega, \end{aligned}$$

where Ω is a bounded domain in $\mathbb{R}^n$, $n \geqq 1$, with smooth boundary $\partial\Omega$ and for $k = 1, \cdots, r$, L_k is a second order uniformly elliptic operator. The coupling coefficients are such that $m_{ij} \geqq 0$, $i \neq j$ and for at least one k, $m_{kk}^{+} \neq 0$. We establish the existence of positive characteristic values with associated positive solutions. We also investigate the multiplicity of such characteristic values and establish bifurcation results for nonlinear perturbations of the linear problem.

Key words. coupled elliptic systems, eigenvalue problems, bifurcation in nonlinear systems

AMS(MOS) subject classifications. Primary 35J55, 35F30, 35J65

1. Introduction.

Consider the eigenvalue problem

$$\begin{aligned} L_k u_k &= \lambda \sum_{i=1}^{r} m_{ki} u_i \quad \text{in } \Omega, \\ u_k &= 0 \qquad\qquad\quad \text{on } \partial\Omega, \end{aligned} \tag{1.1}$$

where Ω is a bounded domain in $\mathbb{R}^n$, $n \geqq 1$, with smooth boundary $\partial\Omega$, and for $k = 1, \cdots, r$

$$L_k = \sum_{i,j=1}^{n} a_{ij}^{k} \frac{\partial^2}{\partial x_i \partial x_j} + \sum_{i=1}^{n} a_i^k \frac{\partial}{\partial x_i} + a_0^k \tag{1.2}$$

is a uniformly elliptic differential operator of second order with coefficients continuous on $\bar{\Omega}$ and $a_0^k(x) \geqq 0$, $x \in \bar{\Omega}$. The coefficients m_{ki}, $1 \leqq k$, $i \leqq r$ are also assumed to belong to $C^0(\bar{\Omega}, \mathbb{R})$. The parameter $\lambda \in \mathbb{R}$ is assumed to be positive.

In a recent paper, P. Hess [11] showed that if $m_{ij} \geqq 0$, $i \neq j$ and if for at least one k, $m_{kk}^{+} \neq 0$, then (1.1) has a positive characteristic value with associated nontrivial solution $u = \operatorname{col}(u_1, \cdots, u_r) \in K = \{ v \in C^0(\bar{\Omega}, \mathbb{R}^r)\colon\ v_i \geqq 0,\ 1 \leqq i \leqq r \}$. The purpose of this paper is to examine this important result more closely. We obtain a somewhat more detailed understanding of the multiplicity and character of the nontrivial solutions to (1.1), leading to results on bifurcation questions for associated nonlinear eigenvalue problems.

* Received by the editors April 24, 1984, and in revised form February 5, 1985. This research was supported in part by the National Science Foundation under grant MCS 8121951.

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

‡ Department of Mathematics, University of Utah, Salt Lake City, Utah 84112.

To this end, we shall begin by proving the main result of Hess [11] in a slightly different way, relying on ideas employed earlier by the second author in [18]. We then give an extension of the result to the multiparameter problem

$$\begin{aligned} L_k u_k &= \lambda_k \sum_{i=1}^{r} m_{ki} u_i \quad \text{in } \Omega, \\ u_k &= 0 \qquad\qquad\qquad\quad \text{on } \partial\Omega, \quad k = 1, \cdots, r. \end{aligned} \tag{1.3}$$

Together with some conditions for the uniqueness and simplicity of such characteristic values of (1.1), our result on (1.3) explains how multiplicities greater than one occur in a number of cases. Finally, we apply these results to the problem of positive solutions to nonlinear eigenvalue problems, including the problem of coexistence steady states in the Volterra–Lotka competition model with diffusion, recently studied by Cosner and Lazer [8].

2. Main results. Let L_k, $1 \leqq k \leqq r$ also denote the realization of L_k in $C_0^0(\bar{\Omega}, \mathbb{R})$ subject to Dirichlet boundary conditions. Then L_k: $C_0^0(\bar{\Omega}, \mathbb{R}) \supset \operatorname{dom}(L_k) \to C_0^0(\bar{\Omega}, \mathbb{R})$ is invertible, with compact inverse. Furthermore, L_k^{-1} is a positive operator with respect to the cone of nonnegative functions. Denote by M the matrix $M = (m_{ij})$, $1 \leqq i, j \leqq r$ and think of M as a multiplication operator. (Recall that $m_{ij} \geqq 0$ if $i \neq j$.) Then (1.1) may be written as

$$Lu = \lambda M u, \tag{2.1}$$

where

$$L = \begin{bmatrix} L_1 & & 0 \\ & \ddots & \\ 0 & & L_r \end{bmatrix}, \qquad u = \begin{bmatrix} u_1 \\ \vdots \\ u_r \end{bmatrix},$$

and L^{-1} is a compact operator on $C_0^0(\bar{\Omega}, \mathbb{R}^r)$ which is positive with respect to the cone K in $C_0^0(\bar{\Omega}, \mathbb{R}^r)$, i.e. $L^{-1}(K) \subset K$.

Let us choose $\mu > 0$ such that all elements on the main diagonal of $M + \mu I = M + \mu$, where I is the $r \times r$ identity matix, are positive. Then (2.1) is equivalent to

$$(L + \lambda\mu) u = \lambda (M + \mu) u, \tag{2.2}$$

and since μ and λ are positive $(L + \mu\lambda)^{-1}$ is also a compact operator positive with respect to K. Thus (2.2) is equivalent to

$$u = \lambda (L + \lambda\mu)^{-1} (M + \mu) u. \tag{2.3}$$

We let $A_\lambda = \lambda (L + \lambda\mu)^{-1} (M + \mu)$. The following result then holds.

LEMMA 2.1. *Let $r(A_\lambda)$ denote the spectral radius of A_λ. Then the mapping $\lambda \to r(A_\lambda)$ is continuous on $(0, \infty)$ with $\lim_{\lambda \to 0^+} r(A_\lambda) = 0$.*

Proof. The map A_λ depends continuously on λ in the strong operator topology. Since the family $\{A_\lambda\}$ is a compact family, it follows from a result of Nussbaum [13] that the map $\lambda \to r(A_\lambda)$ is continuous.

LEMMA 2.2. *Assume $m_{kk}^+ \neq 0$ for some $k \in \{1, \cdots, r\}$. Then there exists $\lambda > 0$ such that $r(A_\lambda) \geqq 1$.*

Proof. According to Hess–Kato [12], there exists $\lambda > 0$ and $u_k \in C_0^0(\bar{\Omega}, \mathbb{R})$, $u_k(x) > 0$, $x \in \Omega$ such that

$$L_k u_k = \lambda m_{kk} u_k.$$

Letting $u=\operatorname{col}(0,\cdots,u_k,0,\cdots,0)$ one gets $Lu\leqq\lambda Mu$. Thus $(L+\lambda\mu)u\leqq\lambda(M+\mu)u$. Hence for this value of λ

$$u\leqq\lambda(L+\lambda\mu)^{-1}(M+\mu)u,$$

i.e. $u\leqq A_\lambda u$.

Iterating this inequality, we get $u\leqq A_\lambda^n u$, and since the C_0^0 norm is monotone with respect to the cone K, we get

$$|u|\leqq|A_\lambda^n||u|.$$

Hence $1\leqq|A_\lambda^n|^{1/n}$. Thus $r(A_\lambda)\geqq 1$.

THEOREM 2.3. *Let $m_{kk}^+\neq 0$ for some $k\in\{1,2,\cdots,r\}$. Then there exists a smallest $\bar\lambda>0$ and $u\in K\setminus\{0\}$ such that*

$$u=A_{\bar\lambda}u,$$

i.e. $Lu=\bar\lambda Mu$.

Proof. Since $r(A_\lambda)$ is continuous and $r(A_\lambda)\to 0$ as $\lambda\to 0$ and since by Lemma 2.2, there exists λ such that $r(A_\lambda)\geqq 1$, it follows that there exists a smallest $\bar\lambda$ such that

$$r(A_{\bar\lambda})=1.$$

Since $A_{\bar\lambda}$ is positive and compact and K is total one may employ the theorem of Krein–Rutman (see [2]) to conclude that there exists $u\in K\setminus\{0\}$ such that

$$u=r(A_{\bar\lambda})u=A_{\bar\lambda}u,$$

i.e., $Lu=\bar\lambda Mu$.

COROLLARY 2.4. *If $\lambda>0$ is any other characteristic value, then $\lambda\geqq\bar\lambda$.*

Proof. Let λ be a characteristic value. Then there is $u\neq 0$ such that $u=A_\lambda u$. Iterating, one obtains $u=A_\lambda^n u$. Hence $1\leqq|A_\lambda^n|^{1/n}$, implying that $r(A_\lambda)\geqq 1$. The result then follows from the proof of Theorem 2.3.

Now consider (1.3). Assume that $(\lambda_1,\cdots,\lambda_r)$ is restricted to a ray emanating from the origin of $\mathbb{R}^r$ into the positive cone. Theorem 2.3 then obtains in most cases. To see this, observe that if $(\lambda_1,\cdots,\lambda_r)$ is as restricted, (1.3) is equivalent to

$$Lu=t\tilde M u, \tag{2.4}$$

where

$$\tilde M=\begin{bmatrix}\lambda_1^0 & & \\ & \ddots & \\ & & \lambda_r^0\end{bmatrix}M,\qquad t\in I,$$

$\lambda_1^0\geqq 0$ and $(\lambda_1^0)^2+\cdots+(\lambda_r^0)=1$. The result follows provided $\tilde m_{kk}^+=\lambda_k^0 m_{kk}^+\neq 0$ for some $k\in\{1,2,\cdots,r\}$.

Suppose now that $\lambda=(\lambda_1,\cdots,\lambda_r)$ is such that $\lambda_i\geqq 0$ and $|\lambda|^2>0$. Define

$$A_\lambda=|\lambda|^2\left(L+|\lambda|^2\mu\right)^{-1}(\tilde M+\mu)$$

with

$$\tilde M=\frac{1}{|\lambda|^2}\begin{bmatrix}\lambda_1 & & \\ & \ddots & \\ & & \lambda_r\end{bmatrix}M.$$

If $|\lambda^0|^2=1$, and $\lambda_i^0>0$, $i=1,\cdots,r$, we define $F(\lambda^0)=\bar{t}(\lambda^0)$, where $\bar{t}>0$ is the smallest number such that $r(A_{\bar{t}\lambda^0})=1$. There are a number of conditions under which the function F is continuous. In particular, we have the following result.

THEOREM 2.5. *F is continuous at $\lambda^0\in(\mathbb{R}_+)^r\cap S^{r-1}$ provided any one of the following conditions holds*:

i) *$m_{ij}\geqq 0$ for $i,j=1,\cdots,k$.*

ii) *If $F(\lambda^0)=\bar{t}\lambda^0$, then $\bar{t}$ is the only positive number t such that $u=A_{t\lambda^0}u$ has a nontrivial solution $u\in K$.*

Proof. In case (i), $r(A_{t\lambda^0})$ is a nondecreasing function of t, $t\geqq 0$. The result follows from an application of a result of Nussbaum [13]. In case (ii), one uses a simple compactness argument.

3. Coupling and multiplicity. In this section we investigate the question of uniqueness and multiplicity of the characteristic values of (1.1). We begin with the following lemma.

LEMMA 3.1. *Suppose $\mu\geqq 0$ is chosen so that all entries of the matrix $M+\mu$ are nonnegative and that there exists $x_0\in\Omega$ such that $(M+\mu)(x_0)$ is irreducible. Then if $Lu=\lambda Mu$ and $u\in K\setminus\{0\}$, then in fact $u\in\operatorname{int}K$ (where* int *is with respect to the $C^1(\bar{\Omega})$ topology).*

Proof. Since $Lu=\lambda Mu$ we have that

$$(L+\lambda\mu)u=\lambda(M+\mu)u,$$

and componentwise

$$(L_i+\lambda\mu)u_i=\lambda\sum_{j=1}^{r}(m_{ij}+\mu\delta_{ij})u_j\geqq 0.$$

It then follows from the Hopf maximum principle that for each $i\in\{1,\cdots,r\}$, either $u_i\equiv 0$ on $\bar{\Omega}$ or $u_i(x)>0$, for all $x\in\Omega$.

Since $u\in K\setminus\{0\}$, there must be at least one $i_0\in\{1,\cdots,r\}$ such that $u_{i_0}(x)>0$ on Ω. If there is no other such i, $(M+\mu)(x_0)$ has a one-dimensional invariant subspace, contradicting the assumption of irreducibility. Hence there is $i_1\in\{1,\cdots,r\}$, $i_1\neq i_0$ such that $u_{i_1}(x)>0$ on Ω. If i_0 and i_1 are the only such, then $(M+\mu)(x_0)$ has a two-dimensional invariant subspace, again a contradiction. Iterating this argument now guarantees $u_i(x)>0$ on Ω for $i=1,\cdots,r$. It further follows from [10, Lemma 3.4] that $\partial u_i/\partial\nu$ is negative on $\partial\Omega$ for each i, proving the result.

THEOREM 3.2. *Suppose there are $\mu\geqq 0$ and $x_0\in\Omega$ such that $(M+\mu)(x_0)$ is irreducible. Then if $\bar{\lambda}$ is as in Theorem* 2.3, *$\bar{\lambda}$ is a geometrically simple eigenvalue.*

Proof. Suppose u and v are nontrivial solutions of (1.1) with $u\in K$. Lemma 3.1 implies $u\in\operatorname{int}K$. Hence for small δ, $u-\delta v\in\operatorname{int}K$. Let $\delta^*=\sup\{\delta>0\colon u-\delta v\in\operatorname{int}K\}$, $\delta_*=\inf\{\delta<0\colon u-\delta v\in\operatorname{int}K\}$. Since $v\neq 0$ at least one of δ^* or δ_* must be finite. With no loss of generality, assume $\delta^*<\infty$. Then $u-\delta^*v\in\partial K$. Lemma 3.1 implies $u\equiv\delta^*v$.

LEMMA 3.3. *Suppose there are $\mu\geqq 0$ and $x_0\in\Omega$ such that $(M+\mu)(x_0)$ is irreducible. Then if $\bar{\lambda}$ is as in Theorem* 2.3, *$(I-A_{\bar{\lambda}})^2z=0$ implies $(I-A_{\bar{\lambda}})z=0$.*

Proof. Suppose $(I-A_{\bar{\lambda}})^2z=0$. Then Theorem 3.2 implies that $(I-A_{\bar{\lambda}})z=cu$, where $u=A_{\bar{\lambda}}u$. Let $A_{\bar{\lambda}}^*$ denote the Banach space adjoint of $A_{\bar{\lambda}}$ considered in $C_0(\bar{\Omega},\mathbb{R}^r)$. The Krein–Rutman theorem implies there is a continuous linear functional f^* (with $f^*(K)\subset[0,\infty]$ and $f^*(\operatorname{int}K)\subset(0,\infty)$) such that $A_{\bar{\lambda}}^*f^*=f^*$. Hence

$$f^*z-f^*A_{\bar{\lambda}}z=cf^*(u)$$

which implies $0=cf^*(u)$. Hence, since $u\in\operatorname{int}K$, $c=0$.

THEOREM 3.4. *Suppose* $m_{ii} \geqq 0$ *for* $i = 1, \cdots, r$. *Then if* $M(x_0)$ *is irreducible for some* $x_0 \in \Omega$ *and* $\bar{\lambda}$ *is as in Theorem* 2.3, $\bar{\lambda}$ *is an algebraically simple eigenvalue and the only positive eigenvalue admitting a solution* u, *where* $u \in K$.

Proof. That $\bar{\lambda}$ is an algebraically simple eigenvalue is a consequence of Lemma 3.3. Suppose now $\lambda' > \bar{\lambda}$ is such that $v = \lambda' L^{-1} M v$, with $v \in K$. Then Lemma 3.1 implies $v \in \operatorname{int} K$. Furthermore, since $r(A_\lambda)$ is a strictly increasing function of λ, $r(A_{\lambda'}) > 1$. The Krein–Rutman theorem implies the existence of $\tilde{v} \in K \setminus \{0\}$ such that $r(A_{\lambda'})\tilde{v} = A_{\lambda'}\tilde{v}$. Applying Lemma 3.1 to the equation

$$L\tilde{v} = \frac{\lambda'}{r(A_{\lambda'})} M\tilde{v},$$

we conclude that in fact $\tilde{v} \in \operatorname{int} K$. For $\delta > 0$ and sufficiently small, $v - \delta\tilde{v} \in \operatorname{int} K$. Let $\delta^* = \sup\{\delta > 0 \colon v - \delta\tilde{v} \in \operatorname{int} K\}$. Since $\tilde{v} \in \operatorname{int} K$, $\delta^* < \infty$ and $v - \delta^*\tilde{v} \in K$, i.e., $\delta^*\tilde{v} \leqq v$. Hence $\delta^* r(A_{\lambda'})\tilde{v} = \delta^* A_{\lambda'} v \leqq A_{\lambda'} v = v$. It follows that $\delta^* r(A_{\lambda'}) \leqq \delta^*$, and so $r(A_{\lambda'}) \leqq 1$, a contradiction.

We do not know in general whether Theorem 3.4 remains valid if the assumption $m_{ii} \geqq 0$ for $i = 1, \cdots, r$ is removed. However, with some additional restrictions on the system (1.1), the theorem remains valid. As we shall see, the restrictions are substantial. Nevertheless, the result is quite useful from the point of view of applications to nonlinear analysis. Before stating the result, we give two lemmas which will be needed in the proof.

LEMMA 3.5. *Suppose that* $m_{ij}(x) \leqq 1/r$, $i \neq j$, *and that* $-1 \leqq m_{ii} \leqq -1 + 1/r$ *for* $j, i = 1, \cdots, r$. *Then there is no positive eigenvalue for* (1.1) *admitting a solution in* $K \setminus \{0\}$.

Proof. The result follows from an application of the maximum principle. See [14, pp. 188–192].

LEMMA 3.6. *Let* $m_{ii} + 1 > 0$ *for* $i = 1, \cdots, r$, *let* $M + I$ *be irreducible, and assume there is* $\lambda > 0$ *and* $u \in K \setminus \{0\}$ *such that* $Lu = \lambda Mu$. *Then* $N((I - \lambda L^{-1}M)^2) = N(I - \lambda L^{-1}M) = \langle u \rangle$, *whenever* $L^{-1}M = ML^{-1}$.

Proof. That $N(I - \lambda L^{-1}M) = \langle u \rangle$ follows from the fact that $u \in \operatorname{int} K$ (see Lemma 3.1). Consider $\lambda(L + \lambda)^{-1}(M + 1)$. By the proof of Theorem 3.4, $r(\lambda(L + \lambda)^{-1}(M + 1)) = 1$. If A_λ^* denotes the Banach space adjoint of $\lambda(L + \lambda)^{-1}(M + 1)$, the Krein–Rutman theorem guarantees the existence of continuous linear functional f^* such that $A_\lambda^* f^* = f^*$ and such that $f^*(\operatorname{int} K) \subset (0, \infty)$.

Suppose $(L - \lambda M)^2 x = 0$. Then $Lx - \lambda Mx = cu$, for some $c \in \mathbb{R}$. Hence $(L + \lambda)x - \lambda(M + 1)x = cu$, or equivalently, $x - \lambda(L + \lambda)^{-1}(M + 1)x = c(L + \lambda)^{-1}u$. It follows that $f^*x - f^*(\lambda(L + \lambda)^{-1}(M + 1)x) = cf^*((L + \lambda)^{-1}u)$. Now $f^*(\lambda(L + \lambda)^{-1}(M + 1)x) = (A_\lambda^* f^*)x = f^*x$. So $0 = cf^*((L + \lambda)^{-1}u)$. Since $u \in \operatorname{int} K$, $c = 0$.

Finally, if $L^{-1}M = ML^{-1}$ and $(I - \lambda L^{-1}M)^2 x = 0$, then a simple computation shows $(L - \lambda M)^2 x = 0$. Hence $(L - \lambda M)x = 0$, or equivalently, $(I - \lambda L^{-1}M)x = 0$.

Remark 3.7. We note that in Lemma 3.6 that we do not need $m_{ii} > 0$ for some $i \in \{1, 2, \cdots, r\}$.

THEOREM 3.8. *Suppose that the conditions of Theorem* 2.3 *are satisfied. In addition, assume*

(i) $m_{ij} \leqq 1/r$, *if* $i \neq j$;

(ii) $-1/2r < m_{ii} < 1/2r$, *for* $i = 1, \cdots, r$.

If $L^{-1}M = ML^{-1}$, $(M + I)$ *is irreducible, and if* $\bar{\lambda}$ *is as in Theorem* 2.3, *then* $\bar{\lambda}$ *is a simple eigenvalue for* (1.1) *and the only positive eigenvalue admitting a solution* u, *with* $u \in K \setminus \{0\}$.

Remark 3.9. (a) Since (1.1) may be rescaled, conditions (i) and (ii) above are restrictions only on the relative sizes of the diagonal versus off-diagonal terms of the matrix M. The commutativity condition requires that $L_i = L_j$ for $i,j = 1,\cdots,r$, and that M be constant, although M may have negative entries on its main diagonal.

(b) The proof relies on an "unfolding" of the problem in a manner analogous to that employed in [12]. We also obtain partial results in case $(M+I)(x_0)$ is irreducible for some $x_0 \in \Omega$ (dropping the commutativity assumption).

Proof of Theorem 3.8. Assume initially only that M satisfies conditions (i) and (ii) of the hypotheses and that $(M+I)(x_0)$ is irreducible for some $x_0 \in \Omega$. Let $\lambda > 0$ and $t \in \mathbb{R}$ and define $A_{\lambda,t}$ by

$$A_{\lambda,t} = \lambda(L+\lambda)^{-1}(M-t+1).$$

We first observe that there is $t^* \in (0, 1-1/2r)$ such that $A_{\lambda,t}$ is a positive operator for $\lambda > 0$ and $t \leq t^*$ and that the equation

$$u = A_{\lambda,t}u \tag{3.1}$$

has no solution with $\lambda > 0$, $t = t^*$, and $u \in K\setminus\{0\}$. To see that this is the case, consider

$$-\frac{1}{2r} < \min_{x\in\bar{\Omega}} m_{ii}(x) \leqq \max_{x\in\bar{\Omega}} m_{ii}(x) < \frac{1}{2r}$$

for some $i \in \{1,2,\cdots,r\}$. It follows that if $0 < t < 1-1/2r$, then $\min_{x\in\bar{\Omega}} m_{ii}(x) - t > -1$, and that if $t > \max_{x\in\bar{\Omega}} m_{ii}(x) + 1 - \frac{1}{r}$, then $\max_{x\in\bar{\Omega}} -m_{ii}(x) - t \leqq -1 + 1/r$. Since

$$\max_{x\in\bar{\Omega}} m_{ii}(x) + 1 - \frac{1}{r} < \frac{1}{2r} + 1 - \frac{1}{r} = 1 - \frac{1}{2r},$$

our observation follows from Lemma 3.5.

Next observe that if $t \leqq 0$, Theorem 2.3 implies that there exists a smallest positive number $\bar{\lambda}(t)$ such that (3.1) has a solution $u \in K\setminus\{0\}$. If $t = t^*$, no such number exists. We now define a function $f\colon (-\infty, t^*] \to [0,\infty)$ by

$$f(t) = \begin{cases} 1/\bar{\lambda}(t) & \text{provided } \bar{\lambda}(t) \text{ exists,} \\ 0 & \text{otherwise.} \end{cases}$$

Suppose now $t < t' \leqq t^*$ and there exists $\lambda(t') > 0$ and $u \in K\setminus\{0\}$ such that

$$u = \lambda(t')\big(L+\lambda(t')\big)^{-1}(M-t'+1)u.$$

Since $0 < m_{ii} - t' + 1 < m_{ii} - t + 1$ for $i = 1,\cdots,r$, we have

$$u \leqq \lambda(t')\big(L+\lambda(t')\big)^{-1}(M-t+1)u.$$

It follows that $\bar{\lambda}(t)$ exists and $\bar{\lambda}(t) \leqq \bar{\lambda}(t')$. Furthermore, if $f(t') = 0$, $f(t) = 0$ for $t \in [t', t^*]$. Hence f is a monotonic nonincreasing function and, as such, can have at most a countable number of discontinuities.

Suppose now that $\lambda_0 > 0$, $t_0 < t^*$, and $u \in K\setminus\{0\}$ such that

$$u = \lambda_0(L+\lambda_0)^{-1}(M-t_0+1)u.$$

Lemma 3.1 implies that $u \in \operatorname{int} K$. It follows from Lemma 3.3 that 1 is a simple eigenvalue of $\lambda_0(L+\lambda_0)^{-1}(M-t_0+1)$ and that $r(\lambda_0(L+\lambda_0)^{-1}(M-t_0-1)) = 1$. Since $r(\lambda(L+\lambda)^{-1}(M-t+1))$ depends continuously on λ and t, a perturbation theory

argument will show that if $\alpha(\lambda,t)=r(\lambda(L+\lambda)^{-1}(M-t+1))$, α is an analytic function of (λ,t) in a neighborhood of(λ_0,t_0). Furthermore, one may choose an eigenfunction $u(\lambda,t)$ corresponding to $\alpha(\lambda,t)$ so that $u(\lambda,t)$ is analytic in (λ,t) also. (The perturbation theory argument necessary can be adapted from [17, pp. 57–64]. We note only that the simplicity of the eigenvalue 1 of $\lambda_0(L+\lambda_0)^{-1}(M-t_0+1)$ is essential to the argument.) Let us now consider the equation

$$\alpha u=\lambda(L+\lambda)^{-1}(M-t+1)u. \tag{3.2}$$

Differentiating (3.2) with respect to t and evaluating at (λ_0,t_0) we obtain

$$\begin{aligned}&\alpha_t(\lambda_0,t_0)u(\lambda_0,t_0)+u_t(\lambda_0,t_0)\\&\quad=\lambda_0(L+\lambda_0)^{-1}(M-t_0+1)u_t(\lambda_0,t_0)-\lambda_0(L+\lambda_0)^{-1}u(\lambda_0,t_0).\end{aligned} \tag{3.3}$$

Let $A^*_{\lambda_0,t_0}$ be the Banach space adjoint of A_{λ_0,t_0}. The Krein–Rutman Theorem implies that there is a continuous linear functional f^* (with $f^*(\operatorname{int}K)\subset(0,\infty)$) such that $f^*A_{\lambda_0,t_0}=A^*_{\lambda_0,t_0}f^*=f^*$. Applying f^* to (3.3) yields

$$\alpha_t(\lambda_0,t_0)f^*(u(\lambda_0,t_0))=-\lambda_0 f^*\left[(L+\lambda_0)^{-1}u(\lambda_0,t_0)\right]. \tag{3.4}$$

Since $u(\lambda_0,t_0)\in\operatorname{int}K$, $\alpha_t(\lambda_0,t_0)\neq 0$. The Implicit Function Theorem implies the existence of $\delta>0$ and a smooth function g: $(\lambda_0-\delta,\lambda_0+\delta)\to\mathbb{R}$ such that $g(\lambda_0)=t_0$ and $\alpha(\lambda,g(\lambda))=1$.

Since f is nonincreasing, if $t_1<t_2$ and $f(t_1)=f(t_2)>0$, then $f(t)\equiv f(t_1)=1/\bar{\lambda}(t_1)$ for $t\in[t_1,t_2]$. So $\alpha(\bar{\lambda}(t_1),t)=1$ for $t\in[t_1,t_2]$ by Lemma 3.3. But for any $t_0\in(t_1,t_2)$, the preceding argument shows that the solution set to $\alpha(\lambda,t)=1$ is expressible as a function of λ in a neighborhood of $(\bar{\lambda}(t_1),t_0)$, a contradiction. Hence f is strictly decreasing so long as it remains positive.

Let $t^{**}\in(0,1-1/2r)$ be given by $t^{**}=\inf\{t\leqq t^*\colon f(t)=0\}$ and also let $\gamma=\lim_{t\to-\infty}f(t)$ and $0\leqq\omega=\inf\{f(t)\colon f(t)>0\}$. Since f is strictly decreasing, it has an inverse h defined from a subset of (ω,γ) into $(-\infty,t^{**})$. We claim that this function h is extendable to a continuous function $\tilde{h}$: $(0,\gamma)\to(-\infty,t^{**})$ such that if $s\in(0,\gamma)$, $\alpha(1/s,\tilde{h}(s))=1$.

Let us now establish this claim. Let $t_0<t^{**}$ and let

$$L_0=\lim_{t\to t_0^-}f(t)\geqq\lim_{t\to t_0^+}f(t)=R_0>0.$$

It follows from [13] that

$$\alpha\left(\frac{1}{L_0},t_0\right)=1\quad\text{and}\quad\alpha\left(\frac{1}{R_0},t_0\right)=1.$$

The minimality of $\bar{\lambda}(t_0)$ implies that $L_0=f(t_0)$. Notice that if $t\leqq-1/2r$, Theorem 3.4 implies that $L_0=R_0$. So if $L_0>R_0$, $t_0\in(-1/2r,t^{**})$. Furthermore, the Implicit Function Theorem may be applied as before at $(1/L_0,t_0)$ and $(1/R_0,t_0)$, giving functions g_1 and g_2 respectively. Notice that: if $\lambda\in[1/L_0,1/R_0]$ and $g_1(\lambda)$ is defined, then Theorem 3.4 and the minimality of $\bar{\lambda}(t)$ for $t>t_0$ implies that $g_1(\lambda)\in[-1/2r,t_0]$, and similarly for $g_2(\lambda)$. By [13], $g_1(\lambda)$ and $g_2(\lambda)$ can be extended on $[1/L_0,1/R_0]$. Since $A_{\tilde{\lambda},t}$ is monotonic in t for fixed $\tilde{\lambda}$, so must $r(A_{\tilde{\lambda},t})$ be. Hence if $g_1(\lambda)\neq g_2(\lambda)$ for some $\lambda\in[1/L_0,1/R_0]$, $\alpha(\lambda,t)=1$ for t between $g_1(\lambda)$ and $g_2(\lambda)$, a contradiction to the Implicit Function Theorem. Hence $\tilde{h}$ may be defined on $[R_0,L_0]$ by $\tilde{h}(s)=g_1(1/s)$.

We have now shown the existence of $\tilde{h}$ on (ω,γ). If $\omega=0$, there is nothing more to do. If $\omega>0$, then $\alpha(1/\omega,t^{**})=1$ and the Implicit Function Theorem guarantees the existence of a g as before with $g(\lambda)\in[-1/2r,t^{**}]$ and $\alpha(\lambda,g(\lambda))=1$. Since $g(\lambda)\in[-1/2r,t^{**}]$, it again follows from the continuity of the spectral radius [13] that g is extendable to $(1/\omega,\infty)$. Defining $\tilde{h}(s)=g(1/s)$ for $s\in(0,\omega)$ completes the verification of the claim.

Now assume that $L^{-1}M=ML^{-1}$. Suppose there is a $t_0\in(-1/2r,t^{**})$ such that $L_0>R_0$. Now $1/L_0=\bar{\lambda}(t_0)<1/R_0<\bar{\lambda}(t)$ for $t>t_0$. Lemma 3.6 implies there exists a $\delta\in(0,1)$ such that the Leray–Schauder indices $\operatorname{ind}_{LS}(I-(1+\delta)\bar{\lambda}(t_0)L^{-1}(M-t_0))$ and $\operatorname{ind}_{LS}(I-(1-\delta)\bar{\lambda}(t_0)L^{-1}(M-t_0))$ are defined and unequal. We may also assume that $\delta>0$ is sufficiently small so that $((1+\delta)\bar{\lambda}(t_0)<1/R_0$. The homotopy invariance property of the Leray–Schauder degree guarantees that

$$\operatorname{ind}_{LS}\big(I-(1+\delta)\bar{\lambda}(t_0)L^{-1}(M-t)\big)\neq\operatorname{ind}_{LS}\big(I-(1-\delta)\bar{\lambda}(t_0)L^{-1}(M-t)\big)$$

for $t\in(t_0,t_0+\varepsilon)$ for $\varepsilon>0$ and sufficiently small. Hence for $t\in(t_0,t_0+\varepsilon)$ there exist $0<\lambda<\bar{\lambda}(t)$ with $N(I-\lambda L^{-1}(M-t))\neq\{0\}$, a contradiction. Hence $L_0=R_0$ and so f is continuous on $(-\infty,t^{**})$. A similar argument gives $\lim_{t\to t^{**}}f(t)=0$ in this case.

The uniqueness assertion of the theorem is now evident from the Implicit Function Theorem and the monotonicity of $r(A_{\lambda,t})$ in t.

COROLLARY 3.10. *Suppose M satisfies* (i) *and* (ii) *of Theorem* 3.8 *and $(M+I)(x_0)$ is irreducible for some $x_0\in\Omega$. Then if $\tilde{h}$, $A_{\lambda,t}$, and γ are as in the proof of Theorem* 3.8, *then the set* $\{(\lambda,t)\in(0,\infty)\times(-\infty,t^{**}]\colon u=A_{\lambda,t}u \text{ for some } u\in K\setminus\{0\}\}=\{(\lambda,h(\frac{1}{\lambda}))\colon \lambda\in(\frac{1}{\gamma},\infty)\}$.

The requirement that $(M+\mu)(x_0)$ be irreducible for some $x_0\in\Omega$ represents a rather strong coupling in the equations of the system. The other extreme is an uncoupled system, i.e. $m_{ij}\equiv 0$ on Ω if $i\neq j$. Both, however, may be viewed as special cases of the following.

Condition 3.11. There is a finite sequence of row and column interchanges under which M is equivalent to a matrix of the form

$$\begin{bmatrix} M_1 & & \\ & \ddots & \\ & & M_l \end{bmatrix} \tag{3.5}$$

where M_i is an $r_i\times r_i$ matrix, with $r_1+\cdots+r_l=r$. Furthermore, (2.1) is equivalent (under an appropriate relabelling of the u_i's) to the collection of systems

$$L_iv_i=\lambda M_iv_i,\quad i=1,\cdots,l \tag{3.6}$$

where

$$v_i=\begin{bmatrix} w_1 \\ \vdots \\ w_{r_i} \end{bmatrix}$$

has the property that $v_i\in\partial K_i$ and v_i a solution of (3.6) implies $v_i\equiv 0$.

As is evident, not all systems of the form (1.1) satisfy Condition 3.11. However, there are sufficient conditions other than the above special cases under which Condition 3.11 holds. We will not dwell on these here.

Suppose now Condition 3.11 is satisfied. Let k_i, $i=1,\cdots,l$ be defined as follows: $k_1=0$, $k_i=\Sigma_{j<i}r_j$, $i=2,\cdots,l$. Then (1.3) can be equivalently expressed as

$$(3.7)\qquad \begin{bmatrix} L_{k_i+1} & & \\ & \ddots & \\ & & L_{k_i+r_i} \end{bmatrix}\begin{bmatrix} u_{k_i+1} \\ \vdots \\ u_{k_i+r_i} \end{bmatrix} = \begin{bmatrix} \lambda_{k_i+1} & & \\ & \ddots & \\ & & \lambda_{k_i+r_i} \end{bmatrix}\begin{bmatrix} m_{k_i+1,k_i+1} & \cdots & m_{k_i+1,k_i+r_i} \\ \vdots & & \vdots \\ m_{k_i+r_i,k_i+1} & \cdots & m_{k_i+r_i,k_i+r_i} \end{bmatrix}\begin{bmatrix} u_{k_i+1} \\ \vdots \\ u_{k_i+r_i} \end{bmatrix},$$

$i=1,\cdots,l$. Let F_i denote the function of Theorem 2.5 relative to (3.6) (i). We have the following result, which is similar in spirit to the results of [6].

THEOREM 3.12. *Suppose Condition* 3.11 *holds and that for each* $i=1,\cdots,l$, *there exists* $d(i)\in\{1,\cdots,r_i\}$ *such that* $(m_{k_i+d(i),k_i+d(i)})^+\neq 0$. *Assume also that* (ii) *of Theorem* 2.5 *holds for* $\{(\lambda_{k_{i+1}},\cdots,\lambda_{k_i+r_i})\in\mathbb{R}^{r_i}_+\colon |(\lambda_{k_{i+1}},\cdots,\lambda_{k_i+r_i})|=1\}$. *Then the set* $\{(\lambda_1,\cdots,\lambda_r)\in\mathbb{R}^r_+\colon$ (3.6) *has nontrivial solution in* $K\}=\cup_{i=1}^{l}T_i$, *where* $T_i=(\mathbb{R}_+)^{k_i}\times \operatorname{im}(F_i)\times(\mathbb{R}_+)^{r-(k_i+r_i)}$. *Furthermore, the geometric multiplicity of* $(\lambda_1,\cdots,\lambda_r)$ *is precisely the number of sets* T_i *of which* $(\lambda_1,\cdots,\lambda_r)$ *is a member.*

Proof. The result follows easily from Theorem 2.5, the results of §3, and the definition of Condition 3.11.

We conclude this section with a brief examination of the system

$$(3.8)\qquad L_1u=\lambda(u+v),\qquad L_2v=\lambda v.$$

(3.8) is a typical example of a system for which Condition 3.11 fails to hold. Multiplicity results for more general upper-triangular nonsymmetric matrices M may be obtained in a manner analogous to that which follows. To this end, consider

$$(3.9)\qquad L_1u=\lambda(u+v),\qquad L_2v=\mu v.$$

Let λ_1 and λ_2 represent the first eigenvalues for L_1 and L_2, respectively. If (3.8) has a nontrivial cone solution, either $v\equiv 0$ or $\mu=\lambda_2$. In the first case, $\lambda=\lambda_1$. If $\mu=\lambda_2$, however, it follows from the results of [12] that a nontrivial solution $\binom{u}{v}$ with $u>0$ and $v>0$ (note that $u=0$ implies $v=0$) is possible only in case $\lambda<\lambda_1$. In particular, it follows that if $\lambda_1=\lambda_2$, then $\lambda=\lambda_1=\lambda_2$ is a characteristic value of (3.8) which is geometrically but not algebraically simple.

4. Bifurcation results. In [11], Hess combined the result of Theorem 2.3 with the methods of [3] to obtain a bifurcation result for the nonlinear eigenvalue problem

$$(4.1)\qquad Lu=\lambda A(u).$$

Here $A\colon K\to C_0(\bar{\Omega};\mathbb{R}^r)$ is the Nemytskii operator associated with a continuous function $a\colon \bar{\Omega}\times\overline{(\mathbb{R}^+)^r}\to\mathbb{R}^r$. He assumed that a satisfies the following conditions:

$$(4.2)\qquad a(x,0)=0,\qquad x\in\bar{\Omega}.$$

(4.3) There exists an $r\times r$ matrix m of functions $m_{kl}\in C(\bar{\Omega};\mathbb{R})$ such that

$$a(x,\sigma)=m(x)\sigma+o(|\sigma|)$$

as $|\sigma|\to 0$, $\sigma\in(\mathbb{R}^+)^r$ (uniformly for $x\in\bar{\Omega}$).

(4.4) There exists a number $\alpha \geqq 0$ such that

$$a(x,\sigma) \geqq -\alpha\sigma$$

for $(x,\sigma) \in \bar{\Omega} \times (\mathbb{R}^+)^r$.

Under the above conditions, Hess showed that if $\Sigma^+ = \{(\lambda,u) \in \mathbb{R} \times C_0^0(\bar{\Omega},\mathbb{R}^r)$: $\lambda > 0$, $u \in K$, $Lu = \lambda A(u)\}$, Σ^+ contains an unbounded component Σ_0 emanating from $(\lambda^*,0)$, where $\lambda^* \geqq \bar{\lambda}$ ($\bar{\lambda}$ as in Theorem 2.3). He also identified certain cases when $\lambda^* = \bar{\lambda}$ (namely $m_{kl} \geqq 0$ for k, $l = 1,\cdots,r$ or $m_{kl} \equiv 0$ if $k \neq l$).

Our results show that if $(M+\mu)(x_0)$ is irreducible for some $x_0 \in \Omega$, then, at least locally, if $(\lambda,u) \in \Sigma_0$ and $u \neq 0$, then $u \in \operatorname{int} K$. Furthermore, if a is independent of x and $L_i = L_j$ for $i,j = 1,\cdots,r$, Theorem 3.8 implies that also in this case $\lambda^* = \bar{\lambda}$.

We now establish some results on the multidimensionality of the nontrivial bifurcating solutions for nonlinear analogues of (1.3). The principal techniques for establishing such are the theorems of Alexander and Antman [1] and Fitzpatrick, Massabo, and Pejsachowitz [9]. Both results require that one work in an open subset of $\mathbb{R}^r \times E$, E an appropriate Banach space. (In this case, $[C_0^{1,\alpha}(\bar{\Omega},\mathbb{R})]^r$ is suitable, where $0 < \alpha < 1$, provided the coefficients of L_k and the m_{kl} are in $C^\alpha(\bar{\Omega},\mathbb{R})$. A precise definition of the spaces is found in [10].) A formulation based on [3] is not immediate. We therefore consider

$$Lu = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_r \end{bmatrix} Mu + H(\lambda_1,\cdots,\lambda_r,u), \tag{4.5}$$

where H: $R^r \times R^r \to R^r$ is continuous, cone-preserving and $H(\lambda_1,\cdots,\lambda_r,w_1,\cdots,w_r) = o(|(w_1,\cdots,w_r)|)$ (uniformly for $(\lambda_1,\cdots,\lambda_r)$ in compact subsets of R^r). It follows from [5] that the techniques of [1] are applicable to (4.5) provided $(\lambda_1,\cdots,\lambda_r)$ is an algebraically simple characteristic value of (1.3).

THEOREM 4.1. *Suppose there is $k \in \{1,\cdots,r\}$ such that $m_{kk}^+ \neq 0$ and that for $\mu \geqq 0$ and sufficiently large $(M+\mu)(x_0)$ is nonnegative irreducible for some $x_0 \in \Omega$. (If $\mu \geqq 0$ is required, assume that the conditions of Theorem* 3.8 *also hold.) Suppose that $H(\lambda_1,\cdots,\lambda_r,-u) = -H(\lambda_1,\cdots,\lambda_r,u)$ and that F is as in Theorem* 2.5. *Then there emanates from $\operatorname{im} F \times \{0\} \cap (\mathbb{R}_+)^r \times E$ a connected set S of solutions to* (4.5) *such that*

(i) $(\lambda_1,\cdots,\lambda_r,u) \in S$ *implies either* $u \in \operatorname{int}$ *or* $(\lambda_1,\cdots,\lambda_r,u) \in \operatorname{im} F \times \{0\}$;

(ii) $S \setminus (\operatorname{im} F \times \{0\})$ *is of topological dimension at least r at every point.*

Remark 4.2. (i) For a precise definition of topology dimension, see [1] and its references.

(ii) If $(\lambda_1,\cdots,\lambda_r)$ is restricted to a ray emanating from the origin of $\mathbb{R}^r$, the global bifurcation alternatives of Rabinowitz [15] hold. (See also [9].) It then follows that S is unbounded in the sense that $S \cap \partial((\mathbb{R}_+)^r \times C_0(\bar{\Omega},\mathbb{R}^r)) \neq \varnothing$.

(iii) The oddness condition on H is a representative condition guaranteeing the existence of "small" positive solutions. Certainly, other such conditions are possible.

COROLLARY 4.3. *Suppose Condition* 3.11 *holds with $l > 1$. Consider* (4.5) *and assume H of* (4.5) *has the form*

$$H(\lambda_1,\cdots,\lambda_r,u_1,\cdots,u_r) = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_r \end{bmatrix} \tilde{H}(u_1,\cdots,u_r).$$

Consider the problem

$$(4.6)\qquad \begin{bmatrix} L_1 & & \\ & \ddots & \\ & & L_{r_1} \end{bmatrix} \begin{bmatrix} u_1 \\ \vdots \\ u_{r_1} \end{bmatrix} = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_{r_1} \end{bmatrix} M_1 \begin{bmatrix} u_1 \\ \vdots \\ u_{r_1} \end{bmatrix} + \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_{r_1} \end{bmatrix} H^*(u_1, \cdots, u_{r_1}),$$

where

$$H^*(u_1, \cdots, u_{r_1}) = \begin{bmatrix} \tilde{H}_1(u_1, \cdots, u_{r_1}, 0, \cdots, 0) \\ \vdots \\ \tilde{H}_{r_1}(u_1, \cdots, u_{r_1}, 0, \cdots, 0) \end{bmatrix}.$$

Suppose that if $(\lambda_1, \cdots, \lambda_{r_1}, u_1, \cdots, u_{r_1})$ *is a solution of* (4.6) *then* $(\lambda_1, \cdots, \lambda_{r_1}, \lambda_{r_1+1}, \cdots, \lambda_r, u_1, \cdots, u_{r_1}, 0, \cdots, 0)$ *is a solution of* (4.5) *for any choice of* $\lambda_{r_1+1}, \cdots, \lambda_r$. *Then if* M_1, H^* *and* F_1 *satisfy the hypotheses of Theorem* 4.1 *with respect to* (4.6), *there emanates from* $[\operatorname{im} F_1 \times (\mathbb{R}^+)^{r-r_1}] \times \{0\} \subset (\mathbb{R}^+)^r \times [C_0^{1,\alpha}(\bar{\Omega}, \mathbb{R})]^{r_1} \times \{0\}$ *a connected set of cone solutions to* (4.5) *such that* $(\lambda_1, \cdots, \lambda_r, u_1, \cdots, u_{r_1}, 0, \cdots, 0) \in S$ *and* $(u_1, \cdots, u_{r_1}) \neq 0$ *implies* $(u_1, \cdots, u_{r_1}) \in \operatorname{int} K_1$. *Furthermore,* $S \setminus ([\operatorname{im} F_1 \times (\mathbb{R}^+)^{r-r_1}] \times \{0\})$ *has a topological dimension at least r at every point.*

Proof. If $(\lambda_{r_1+1}, \cdots, \lambda_r)$ are considered fixed, Theorem 4.1 guarantees the existence of a set $\tilde{S}$ of solutions to (4.6) as above with topological dimension at least r_i. Then $S = \tilde{S} \times (\mathbb{R}^+)^{r-r_1}$.

Corollary 4.3 raises the obvious question: do there exist other "small" positive solutions with $(u_{r_1+1}, \cdots, u_r$ not identically zero) bifurcating from $[\operatorname{im} F_1 \times (\mathbb{R}^+)^{r-r_1}] \times \{0\}$? The answer is no, provided $(\lambda_1, \cdots, \lambda_r)$ is algebraically simple (with respect to all of M) and H is sufficiently well behaved. More precisely, we have the following result.

COROLLARY 4.4. *Suppose* $(\lambda_1, \cdots, \lambda_r) \in \operatorname{im} F_1 \times (\mathbb{R}^+)^{r-r_1}$ *is as in Corollary* 4.3 *and that*

$$\dim N\left[I - \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_r \end{bmatrix} L^{-1}M\right] = \dim N\left[\left(I - \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_r \end{bmatrix} L^{-1}M\right)^2\right] = 1.$$

Suppose also that there exists a continuous monotonic function $G\colon \mathbb{R}^+ \to \mathbb{R}^+$ *such that* $G(0) = 0$ *and for all* $v, w \in E = [C_0^{1,\alpha}(\bar{\Omega}, \mathbb{R})]^r$

$$\|\tilde{H}(v_1, \cdots, v_r) - \tilde{H}(w_1, \cdots, w_r)\|_E \leqq G\big(\|(v_1, \cdots, v_r)\|_E + \|(w_1, \cdots, w_r)\|_E\big)\|(v_1, \cdots, v_r) - (w_1, \cdots, w_r)\|_E.$$

Then near $(\lambda_1, \cdots, \lambda_r, 0, \cdots, 0)$ *in* $(\mathbb{R}^+)^r \times E$, *the solution set of* (4.5) *is as described by Corollary* 4.3.

Proof. The result follows from the generalization [1, Thm. 3.12] to several parameters of the global bifurcation theorem of Rabinowitz [16, Thm. 1.19].

We conclude this article by demonstrating the applicability of Corollaries 4.3 and 4.4 to the question of stable coexistence states in the Volterra–Lotka competition model with diffusion, recently studied by Cosner and Lazer [8], among others. The model in question is as follows:

$$\begin{aligned} -\Delta u &= au - bu^2 - cuv, \\ -\Delta v &= dv - euv - fv^2, \end{aligned} \tag{4.7}$$

with $u(x)$ and $v(x)$ the population densities of the competing species at $x \in \Omega$, an open bounded smooth domain in $\mathbb{R}^n$, subject to $u \equiv 0 \equiv v$ on $\partial\Omega$. a and d, b and f and c and e are assumed to be positive constants, representing growth rates, self-regularization and competition, respectively, with positive diffusion coefficients normalized to 1. The only solutions to (4.7) with physical significance are those with $u \geqq 0$, $v \geqq 0$.

We will now consider (4.7) as a bifurcation problem with a and d acting as parameters and b, c, e, f considered fixed. Let λ_1 be the first eigenvalue of $-\Delta$ (relative to Ω and zero boundary conditions). Then if F_1 and F_2 are as in Theorem 3.8 relative to

$$-\Delta u = au, \qquad -\Delta v = bv, \tag{4.8}$$

then $\operatorname{im}(F_1) \times R = \{(a,d)\colon a = \lambda_1\}$ and $R \times \operatorname{im}(F_2) = \{(a,d)\colon d = \lambda_1\}$.

Consider the problem

$$\begin{aligned} -\Delta u &= au - bu^2 \quad \text{in } \Omega, \\ u &\equiv 0 \quad \text{on } \partial\Omega. \end{aligned} \tag{4.9}$$

As noted in [8], (4.9) has a unique positive solution for all $a > \lambda_1$. In fact, one may realize these positive solutions as one of the two branches guaranteed by [16, Thm. 1.19]. To see this, note that the nonlinearity in the problem

$$\begin{aligned} -\Delta u &= au - bu|u|, \quad \text{in } \Omega, \\ u &\equiv 0 \quad \text{on } \partial\Omega, \end{aligned} \tag{4.10}$$

is odd. Since the eigenfunction corresponding to λ_1 is of one sign and since $f(x) = x|x|$ is continuously differentiable at $x = 0$, a positive branch for (4.10) is guaranteed at least locally. This branch coincides with solutions to (4.9). One may then use the uniqueness of the positive solutions, upper and lower solution techniques, the maximum principle, and global Rabinowitz bifurcation theory [15] to guarantee the continuation of the branch. Corollary 4.3 and Corollary 4.4 may now be applied. As a result, the only cone solution to (4.7) possible in a sufficiently small neighborhood of $(\lambda_1, d, 0, 0)$, $d \neq \lambda_1$, or of $(a, \lambda_1, 0, 0)$, $a \neq \lambda_1$, are of the form $(a, d, u, 0)$ or $(a, d, 0, v)$, respectively. Thus $a > \lambda_1$, $d > \lambda_1$ does not give a sufficient condition for stable coexistence states if $a \neq d$. This result strongly suggests stable coexistence states should be in general viewed as a secondary bifurcation phenomenon, as is the case in [4] and [7].

Note. It has been observed by one of the referees for this paper that the use of irreducibility in Lemma 3.1 is similar to that in the paper, G. J. Habeitler and M. A. Martino, *Existence theorems and spectral theory for the multigroup diffusion model*, Proc. Symposia in Applied Mathematics XI. Nuclear Reactor Theory, American Mathematical Society, Providence, RI, 1961, pp. 127–139.

REFERENCES

[1] J. C. ALEXANDER AND S. S. ANTMAN, *Global and local behavior of bifurcating multidimensional continua of solutions for multiparameter nonlinear eigenvalue problems*, Arch. Rat. Mech. Anal., 76 (1981), pp. 339–354.

[2] H. AMANN, *Fixed point equations and nonlinear eigenvalue problems in ordered Banach spaces*, SIAM Rev., 18 (1976), pp. 620–709.

[3] _____, *Nonlinear operators in ordered Banach space and some applications to nonlinear boundary value problems*, Lecture Notes in Mathematics 543, Springer, Berlin, 1976, pp. 1–55.

[4] J. BLAT AND K. J. BROWN, *Bifurcation of steady-state solutions in predator-prey and competition systems*, Proc. Roy. Soc. Edinburgh, 97A (1984), pp. 21–34.

[5] R. S. CANTRELL, *A homogeneity condition guaranteeing bifurcation in multiparameter nonlinear eigenvalue problems*, Nonlinear Anal., 8 (1984), pp. 159–169.

[6] _____, *Multiparameter bifurcation problems and topological degree*, J. Differential Equations, to appear.

[7] R. S. CANTRELL AND C. COSNER, *On the steady-state problem for the Volterra–Lotka competition model with diffusion*, Houston J. Math., to appear.

[8] C. COSNER AND A. C. LAZER, *Stable coexistence states in the Volterra–Lotka competition model with diffusion*, SIAM J. Appl. Math., 44 (1984), pp. 1112–1132.

[9] P. M.. FITZPATRICK, I. MASSABO AND J. PEJSACHOWITZ, *Global several-parameter bifurcation and continuation theorems: a unified approach via complementing maps*, preprint.

[10] D. GILBARG AND N. S. TRUDINGER, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, 1977.

[11] P. HESS, *On the eigenvalue problem for weakly-coupled elliptic systems*, Arch. Rat. Mech. Anal., 81 (1983), pp. 151–159.

[12] P. HESS AND T. KATO, *On some linear and nonlinear eigenvalue problems with an indefinite weight function*, Comm. Partial Differential Equations, 5 (1980), pp. 999–1030.

[13] R. NUSSBAUM, *Periodic solutions of some nonlinear integral equations*, Proc. Int. Symp. on Dynamical Systems, A. R. Bednarek and L. Cesari, eds., Gainesville, FL, 1976.

[14] M. H. PROTTER AND H. F. WEINBERGER, *Maximum Principles in Differential Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1967.

[15] P. H. RABINOWITZ, *Some aspects of nonlinear eigenvalue problems*, Rocky Mountain J. Math., 3 (1973), pp. 161–201.

[16] _____, *Some global results for nonlinear eigenvalue problems*, J. Funct. Anal., 77 (1971), pp. 487–513.

[17] F. RELLICH, *Perturbation Theory of Eigenvalue Problems*, Gordon and Breach, New York, 1969.

[18] K. SCHMITT AND H. L. SMITH, *Positive solutions and conjugate points for systems of differential equations*, Nonlinear Anal., 2 (1978), pp. 93–105.

SIAM J. MATH. ANAL.
Vol. 17, No. 4, July 1986

008

A DENSITY DEPENDENT DIFFUSION EQUATION IN POPULATION DYNAMICS: STABILIZATION TO EQUILIBRIUM*

M. BERTSCH† AND D. HILHORST‡

Abstract. We study an evolution problem corresponding to the nonlinear diffusion equation $u_t = \Delta\varphi(u) + \operatorname{div}(u \operatorname{grad} v)$ with no flux boundary conditions. This problem has a continuum of stationary solutions. We prove the existence and uniqueness of the solution of the evolution problem and construct a Lyapunov functional in order to show that the solution stabilizes as $t \to \infty$.

Key words. nonlinear degenerate diffusion equation, large time behavior, Lyapunov functional, population dynamics

AMS(MOS) subject classifications. Primary 35K60, 35K65, 35B40, 34D20

1. Introduction. Let Ω be a bounded domain in $\mathbb{R}^N (N \geqq 1)$ with smooth boundary $\partial\Omega$. We consider the nonlinear evolution problem

$$\text{(P)} \qquad \begin{aligned} &u_t = \Delta\varphi(u) + \operatorname{div}(u \operatorname{grad} v) && \text{in } \Omega \times \mathbb{R}^+, \\ &\frac{\partial}{\partial \nu}\varphi(u) + u\frac{\partial v}{\partial \nu} = 0 && \text{on } \partial\Omega \times \mathbb{R}^+, \\ &u(x,0) = u_0(x) && \text{in } \Omega. \end{aligned}$$

Here ν denotes the outward normal at $x \in \partial\Omega$, the function φ is a smooth function such that $\varphi(0)=0$, $\varphi'(s)>0$ for $s>0$ and $\varphi'(0)=0$, the initial function $u_0 \in L^\infty(\Omega)$ is nonnegative and $v \in W^{1,\infty}(\Omega)$ is a given function (for the precise assumptions we refer to §3).

In §2, we show how Problem P arises in the theory of population dynamics in the case that $\varphi(s) = \frac{1}{2}s^2$ and interpret some of our results in terms of the geographical location of two biological populations.

This paper is divided into two main parts. In Part I (§§4–6) we discuss the large time behaviour of the solution of Problem P. In Part II (§§7–9) we collect the basic results about Problem P: existence, uniqueness and regularity of the solution.

In Part I we prove that the solution $u(t; u_0)$ of Problem P stabilizes to equilibrium. Let E denote the set of equilibrium solutions; then there exists a function $q \in E$ such that

$$u(t; u_0) \to q \quad \text{in } C(\bar{\Omega}) \text{ as } t \to \infty$$

where q satisfies

$$\int_\Omega q\,dx = \int_\Omega u_0\,dx.$$

*Received by the editors May 3, 1983, and in revised form March 21, 1985.

†‡University of Leiden, Leiden, The Netherlands.

‡Present address: CNRS, Laboratoire d'Analyse Numérique, Université de Paris-Sud, 91405 Orsay, France.

In addition we give a characterization of E: we show that E coincides with the set

$$S = \{ w \in C(\bar{\Omega}) : w \geqq 0 \text{ in } \Omega, \text{ and for every } x \in \Omega \text{ either } w(x) = 0 \text{ or } \Phi(w) + v = \text{constant in a neighbourhood of } x \}. \tag{1.1}$$

Here

$$\Phi(s) = \int_0^s \frac{\varphi'(\tau)}{\tau}\, d\tau, \qquad s \geqq 0. \tag{1.2}$$

The proof of these results is given in §§4 and 5. In §4 we show that solutions of Problem P satisfy a contraction property in $L^1(\Omega)$. In §5 we follow an idea of Osher and Ralston [18] and exploit this contraction property, combined with the structure of the set S, to construct a Lyapunov functional. A remarkable detail of the proof is that we do not study the elliptic problem to prove that $E = S$. Also this fact follows from the contraction property and the structure of the set S.

In §6 we extend the above results to the case when the natural boundary condition is replaced by a homogeneous Dirichlet condition.

In Part II, we show that Problem P has a unique solution in some generalized sense. In §7 we construct a solution $u(t; u_0)$ of Problem P as the limit of solutions of related uniformly parabolic problems. It turns out that the set $\{u(t; u_0); t \geqq 1\}$ is precompact in $C(\bar{\Omega})$, thanks to a regularity result of DiBenedetto [7].

In order to show that the solution of Problem P is unique, we are led to use another sequence of regularized problems, following closely a method of Kalashnikov [12], [13]. This is done in §8.

Finally, in §9, we give the corresponding results about the Dirichlet problem.

Studies concerning the existence and uniqueness of the solution of problems related to Problem P have also been done by Aronson, Crandall and Peletier [3], Diaz and Kersner [6], Gagneux [10], Madaune [17] and Touré [21].

There exists an extensive literature about the large time behaviour of solutions of degenerate parabolic equations. However there are not many articles where one constructs a Lyapunov functional in order to establish the stabilization to equilibrium. We have already mentioned the work of Osher and Ralston [18]. Such a method is also used by Aronson, Crandall and Peletier [3], Schatzman [20] and Alikakos and Rostamian [1], [2].

2. Biological context. Problem P arises in the theory of population dynamics. Consider a population in a finite habitat Ω which consists of two different groups, for instance age groups. Let $u(x,t)$ and $v(x,t)$ denote the density of these groups. In order to model their evolution with time, Gurtin and Pipkin [11] propose the following system of equations:

$$u_t = \operatorname{div}(u \operatorname{grad}(u+v)) \quad \text{in } \Omega \times \mathbb{R}^+,$$
$$v_t = k \operatorname{div}(v \operatorname{grad}(u+v)) \quad \text{in } \Omega \times \mathbb{R}^+,$$

where k is some positive constant. The flow of the populations is described by the dispersal velocities: $\operatorname{grad}(u+v)$ for the u-individuals and $k \operatorname{grad}(u+v)$ for the v-individuals. In particular, when the parameter k is small, the v-individuals disperse much slower than the u-individuals.

In this article, we study the problem in the limit $k = 0$. The second equation yields at once that v is constant in time; there remains the equation in u which coincides with the differential equation in Problem P if we set $\varphi(s) = \frac{1}{2}s^2$. The boundary condition expresses the fact that no individuals can leave or enter the habitat.

An interesting consequence of our results is the following. It follows from (1.1) that, for any nonconstant function $v(x)$, the set $E=S$ contains a nontrivial function $q(x)$ such that

$$q\equiv 0 \quad \text{in } \Omega_0\subset\Omega, \qquad q>0 \quad \text{in } \Omega\setminus\Omega_0$$

for some nonempty subset Ω_0. If $u_0\leqq q$ in Ω, then $u(t;u_0)\leqq q$ in Ω for all $t>0$. In particular

$$u(t;u_0)\equiv 0 \quad \text{in } \Omega_0 \text{ for } t\geqq 0.$$

From a biological point of view this phenomenon of *localization* is interesting: the v-individuals can stop the spread of the u-individuals.

A detailed analysis of this model in one space dimension was given in [4].

3. Preliminaries. Let us first state the precise hypotheses on φ, u_0 and v and give a definition of a solution of Problem P.

H1. $\varphi\in C^3(\mathbb{R}^+)\cap C^1(\overline{\mathbb{R}^+})$, $\varphi(0)=\varphi'(0)=0$, $\int_0^1 \tau^{-1}\varphi'(\tau)\,d\tau<\infty$, $\varphi'(s)>0$ for $s>0$, $\varphi''(s)\geqq 0$ for $s\in(0,s_0)$ for some $s_0>0$.

H2a. $v\in W^{1,\infty}(\tilde{\Omega})$ for some smooth domain $\tilde{\Omega}\supset\bar{\Omega}$, and $\Delta v\geqq -M$ in $\tilde{\Omega}$ in the sense of distributions for some $M>0$.

H2b. If $N\geqq 2$, $v\in W^{2,p}(\tilde{\Omega})$ for some $p>N$.

H2c. v has finitely many local strict minima.

H3. If $N=1$ either $\varphi(s)=\frac{1}{2}s^2$ or $v''\in L^1(\Omega)$.

H4. $u_0\in L^\infty(\Omega)$, $u_0\geqq 0$ a.e. in Ω.

We use the notation $Q_t=\Omega\times(0,t]$ for $t>0$ and $Q=\Omega\times\mathbb{R}^+$.

DEFINITION 3.1. We say that a function $u\colon [0,\infty)\to L^1(\Omega)$ is a *generalized solution* of Problem P if it satisfies:

(i) $u\in C([0,t];L^1(\Omega))\cap L^\infty(Q_t)$ for all $t>0$,

(ii) $\int_\Omega u(t)\psi(t)=\int_\Omega u_0\psi(0)+\int\int_{Q_t}\{\varphi(u)\Delta\psi+u\psi_t-u\operatorname{grad}v\operatorname{grad}\psi\}$ for all $t>0$ and all $\psi\in C^{2,1}(\bar{Q})$ such that $\psi\geqq 0$ in Q and $\partial\psi/\partial\nu=0$ on $\partial\Omega\times\mathbb{R}^+$.

A generalized subsolution (resp. supersolution) of Problem P is defined by (i) and (ii) with equality replaced by $\leqq$ (resp. $\geqq$).

In the sequel we shall often omit the word generalized.

In Part II, we shall prove the following results. We suppose that the hypotheses H1, H2a and and H4 are satisfied.

PROPOSITION 3.2. *There exists a unique solution of Problem* P.

PROPOSITION 3.3. (regularity). *Let u be the solution of Problem* P. *Then $u\in C(\bar{\Omega}\times(0,\infty))$ and the set $\{u(t);t\geqq 1\}$ is bounded and equicontinuous. Furthermore if $u_0\in C(\bar{\Omega})$, then $u\in C(\bar{\Omega}\times[0,\infty))$.*

PROPOSITION 3.4. (comparison principle). *Let $\underline{u}(t)$ and $\bar{u}(t)$ be respectively a subsolution and a supersolution of Problem* P *with initial functions $\underline{u}_0$ and $\bar{u}_0$ such that $\underline{u}_0\leqq\bar{u}_0$. Then $\underline{u}(t)\leqq\bar{u}(t)$ in Ω for $t\geqq 0$.*

Part I

4. Contraction in $L^1(\Omega)$. In this section we prove a contraction theorem which turns out to be our main tool when studying the asymptotic behaviour of $u(t)$ as $t\to\infty$.

THEOREM 4.1. *Let $u_1(t)$ and $u_2(t)$ be the solutions of Problem* P *with initial functions u_{01} and u_{02} respectively and suppose that the hypotheses* H1, H2a *and* H4 *are satisfied.*

(i) *Then*

$$\|u_1(t)-u_2(t)\|_{L^1(\Omega)} \leqq \|u_{01}-u_{02}\|_{L^1(\Omega)} \quad \text{for any } t>0.$$

(ii) *Let v satisfy in addition the hypotheses* H2b *and* H3. *If u_{01} and $u_{02} \in C(\bar{\Omega})$ and if there exists a connected subdomain $U \subset \Omega$ such that*

$$\text{(4.1)} \quad \text{and} \quad \begin{aligned} &u_{01}>0, \quad u_{02}>0 \quad \text{in } \bar{U} \\ &u_{01}-u_{02} \text{ changes sign in } U, \end{aligned}$$

then

$$\|u_1(t)-u_2(t)\|_{L^1(\Omega)} < \|u_{01}-u_{02}\|_{L^1(\Omega)} \quad \text{for any } t>0.$$

Remark 4.2. Condition (4.1) is necessary because the parabolic equation in Problem P is degenerate at points where $u=0$.

Due to the degeneracy of the equation and the fact that v is not smooth, the proof of Theorem 4.1 is fairly technical. The idea of the proof is due to Osher and Ralston [18].

Proof of (i). In Part II of this article we show that we can approximate $u_i (i=1,2)$ by solutions of uniformly parabolic problems: let $u_{i\varepsilon} (\varepsilon>0)$ be the classical solution of the problem

$$\begin{aligned} &u_t = \Delta \varphi_\varepsilon(u) + \operatorname{div}(u \operatorname{grad} v_\varepsilon) && \text{in } \Omega \times \mathbb{R}^+, \\ &\frac{\partial}{\partial \nu} \varphi_\varepsilon(u) + u \frac{\partial v_\varepsilon}{\partial \nu} = 0 && \text{on } \partial\Omega \times \mathbb{R}^+, \\ &u(x,0) = u_{0i\varepsilon}(x) && \text{in } \Omega, \end{aligned}$$

where φ_ε is a smooth function such that $\varphi_\varepsilon'(s) \geqq c(\varepsilon)>0$ for $s \geqq 0$ and $\varphi_\varepsilon(s) \to \varphi(s)$ uniformly on compact subsets of $[0,\infty)$ and where v_ε and $u_{0i\varepsilon}$ are smooth functions such that $v_\varepsilon \to v$ in $H^1(\Omega)$ and $u_{0i\varepsilon} \to u_{0i}$ in $L^2(\Omega)$ as $\varepsilon \downarrow 0$. In part II we show that $\{u_{i\varepsilon}\}$ is uniformly bounded and equicontinuous in compact subsets of $\bar{\Omega} \times (0,\infty)$. Using in addition the uniqueness of the solution $u_i (i=1,2)$, we conclude that

$$\text{(4.2)} \qquad u_{i\varepsilon}(t) \to u_i(t) \quad \text{in } C(\bar{\Omega}) \text{ as } \varepsilon \to 0 \text{ for } t>0, \quad i=1,2.$$

We define

$$z_\varepsilon(x,t) = u_{1\varepsilon}(x,t) - u_{2\varepsilon}(x,t), \qquad x \in \bar{\Omega}, \quad t \geqq 0.$$

Then z_ε is the solution of the linear problem

$$\text{(L}_\varepsilon\text{)} \qquad \begin{aligned} &z_t = \Delta(a_\varepsilon z) + \operatorname{div}(z \operatorname{grad} v_\varepsilon) && \text{in } \Omega \times \mathbb{R}^+, \\ &\frac{\partial}{\partial \nu}(a_\varepsilon z) + z \frac{\partial v_\varepsilon}{\partial \nu} = 0 && \text{on } \partial\Omega \times \mathbb{R}^+, \\ &z(x,0) = z_{0\varepsilon}(x) \equiv u_{01\varepsilon}(x) - u_{02\varepsilon}(x) && \text{in } \Omega, \end{aligned}$$

where

$$a_\varepsilon(x,t) = \int_0^1 \varphi_\varepsilon'(\theta u_{1\varepsilon}(x,t) + (1-\theta) u_{2\varepsilon}(x,t))\, d\theta.$$

For smooth initial functions $z_{0\varepsilon}$ which satisfy the compatibility conditions at $\partial\Omega \times \{0\}$, the existence of a unique solution $z_\varepsilon \in C^{2,1}(\bar{Q})$ of Problem L_ε is proved in [16, Thm. 5.3, p. 320]. Below we shall need an existence and uniqueness result if $z_{0\varepsilon}$ is

merely continuous in $\bar{\Omega}$. To obtain this result we can proceed in the same way as we sketched above (and as we shall prove in §7) for the more difficult nonlinear and degenerate Problem P: we approximate $z_{0\varepsilon}$ uniformly by smooth initial functions $z_{0\varepsilon n}$ $(n=1,2,\cdots)$. Then the corresponding solutions $z_{\varepsilon n}$ are uniformly bounded and equicontinuous in $\bar{\Omega}\times[0,t]$ for $t>0$ and $z_{\varepsilon n}$ converges uniformly to a generalized solution $z_\varepsilon \in C(\bar{\Omega}\times[0,t])$ of Problem L_ε as $n\to\infty$. By standard regularity results [16], [8], $z_\varepsilon \in C^{2,1}(\Omega\times(0,t])$. In addition these solutions satisfy the comparison principle; in particular they are uniquely determined by the initial function. The proof rests on the same test function argument which is used in §8 for the nonlinear problem and which is extremely easy in this linear case.

For any initial function $z_0\in C(\bar{\Omega})$, we denote the unique solution of Problem L_ε by $z_\varepsilon(t)=T_\varepsilon(t)z_0$. We set $a^+=\max\{a,0\}$ and $a^-=\max\{-a,0\}$. Then for any $t>0$

$$\begin{aligned}
\|z_\varepsilon(t)\|_{L^1(\Omega)}-\|z_{0\varepsilon}\|_{L^1(\Omega)} &= \|T_\varepsilon(t)z_{0\varepsilon}^+ - T_\varepsilon(t)z_{0\varepsilon}^-\|_{L^1(\Omega)} - \|z_{0\varepsilon}\|_{L^1(\Omega)} \\
&= \int_\Omega \Big\{ \max_{+,-} T_\varepsilon(t)z_{0\varepsilon}^\pm - \min_{+,-} T_\varepsilon(t)z_{0\varepsilon}^\pm \Big\}\, dx - \|z_{0\varepsilon}\|_{L^1(\Omega)} \\
&= \int_\Omega \Big\{ \max_{+,-} T_\varepsilon(t)z_{0\varepsilon}^\pm + \min_{+,-} T_\varepsilon(t)z_{0\varepsilon}^\pm \Big\}\, dx - \int_\Omega (z_{0\varepsilon}^+ + z_{0\varepsilon}^-)\, dx \\
-2\int_\Omega \min_{+,-} T_\varepsilon(t)z_{0\varepsilon}^\pm\, dx &= \int_\Omega \{T_\varepsilon(t)z_{0\varepsilon}^+ + T_\varepsilon(t)z_{0\varepsilon}^- - z_{0\varepsilon}^+ - z_{0\varepsilon}^-\}\, dx \\
-2\int_\Omega \min_{+,-} T_\varepsilon(t)z_{0\varepsilon}^\pm\, dx &= -2\int_\Omega \min_{+,-} T_\varepsilon(t)z_{0\varepsilon}^\pm\, dx,
\end{aligned}$$

since

$$\int_\Omega T_\varepsilon(t)z_{0\varepsilon}^\pm\, dx = \int_\Omega z_{0\varepsilon}^\pm\, dx.$$

It follows from the comparison principle that $T_\varepsilon(t)z_{0\varepsilon}^\pm \geqq 0$. Thus for any $\varepsilon>0$

$$\|u_{1\varepsilon}(t)-u_{2\varepsilon}(t)\|_{L^1(\Omega)} - \|u_{01\varepsilon}-u_{02\varepsilon}\|_{L^1(\Omega)} \leqq 0. \tag{4.3}$$

Clearly Theorem 4.1 (i) follows from (4.2) and (4.3).

Proof of Theorem 4.1(ii). Let $u_{i\varepsilon}(i=1,2)$ and z_ε be defined as above. Since $u_{0i}\in C(\bar{\Omega})$ we may assume that $u_{0i\varepsilon}=u_{0i}$. Let $\delta>0$, and write

$$z_\varepsilon(t)=T_\varepsilon^\delta(t)z_\varepsilon(\delta) \quad \text{for } t\geqq\delta.$$

Then, by the proof of (i),

$$\|z_\varepsilon(t)\|_{L^1(\Omega)} - \|z_\varepsilon(\delta)\|_{L^1(\Omega)} = -2\int_\Omega \min_{+,-}\big(T_\varepsilon^\delta(t)z_\varepsilon^\pm(\delta)\big), \qquad t\geqq\delta,$$

and it is enough to prove that, for sufficiently small values of δ, there exists a $t_1=t_1(\delta)>\delta$ such that

$$\int_\Omega \min_{+,-}\big(T_\varepsilon^\delta(t)z_\varepsilon^\pm(\delta)\big) \geqq \eta(t,\delta)>0 \quad \text{for } t\in(\delta,t_1)$$

for all small $\varepsilon>0$.

Consider the Cauchy–Dirichlet problem

$$(\tilde{\mathrm{L}}_\varepsilon^\delta)\qquad \begin{aligned} &z_t=\Delta(a_\varepsilon z)+\operatorname{div}(z\operatorname{grad} v_\varepsilon) && \text{in } \tilde{U}\times(\delta,\infty),\\ &z=0 && \text{on } \partial\tilde{U}\times(\delta,\infty),\\ &z(\cdot,\delta)=z_\varepsilon^{\pm}(\delta) && \text{in } \tilde{U}, \end{aligned}$$

where $\tilde{U}\subset U$ is such that $\operatorname{dist}(\tilde{U},\delta U)>0$ and $z_{0\varepsilon}=u_{01}-u_{02}$ changes sign in $\tilde{U}$. We denote the solution of Problem $\tilde{L}_\varepsilon^\delta$ by

$$\tilde{z}_\varepsilon^{\pm}(t)=\tilde{T}_\varepsilon^\delta(t)z_\varepsilon^{\pm}(\delta)\quad \text{in } \tilde{U}\times(\delta,\infty).$$

Then, by the maximum principle,

$$\tilde{T}_\varepsilon^\delta(t)z_\varepsilon^{\pm}(\delta)\leqq T_\varepsilon^\delta(t)z_\varepsilon^{\pm}(\delta)\quad \text{in } \tilde{U}\times(\delta,\infty).$$

Thus, it is enough to prove that

$$\int_{\tilde{U}} \min_{+,-}\left(\tilde{T}_\varepsilon^\delta(t)z_\varepsilon^{\pm}(\delta)\right)\geqq\eta(t,\delta)>0\quad \text{for } t\in(\delta,t_1) \tag{4.4}$$

for every $\varepsilon\in(0,\varepsilon_0)$ for some $\varepsilon_0>0$.

This will be done by means of the following lemma, which is an immediate consequence of Harnack's inequality [16, p. 209–210].

LEMMA 4.3. *Let $\varepsilon_0>0$ and $t_1>\delta>0$ be constants, and let the following assumptions be satisfied for all $\varepsilon\in(0,\varepsilon_0)$:*

(a) $\|\tilde{T}_\varepsilon^\delta(t)z_\varepsilon^{\pm}(\delta)\|_{L^\infty(\tilde{U})}\geqq\mu_0>0$ *for* $t\in[\delta,t_1]$,

(b) *when* $N=1$, *then* $\|a_\varepsilon\|_{L^\infty(\delta,t_1;H^1(U))}\leqq C$ *and* $\|v_\varepsilon\|_{W^{1,\infty}(U)}\leqq C$,

(c) *when* $N\geqq 2$, *then* $\|a_\varepsilon\|_{L^\infty(\delta,t_1;W^{1,\infty}(U))}\leqq C$ *and* $\|v_\varepsilon\|_{W^{2,p}(U)}\leqq C$,

(d) $a_\varepsilon\geqq\alpha_0>0$ *in* $\bar{U}\times[\delta,t_1]$, *for some constants* $\mu_0>0$, $C>0$, $\alpha_0>0$ *and* $p>N$.

Then, for all $\varepsilon\in(0,\varepsilon_0)$

$$\tilde{T}_\varepsilon^\delta(t)z_\varepsilon^{\pm}(\delta)\geqq\mu(x,t;\delta)>0\quad \text{in } \tilde{U}\times(\delta,t_1]$$

for some function μ which does not depend on ε.

Assuming that (a), (b), (c) and (d) are satisfied for small $\delta>0$, (4.4) follows. Thus, to complete the proof we need to verify these conditions.

In view of Proposition 3.3 and the assumption $u_{0i}\in C(\bar{\Omega})$, we have $u_i\in C(\bar{\Omega}\times[0,\infty))$. Hence there exists a $t_0>0$ such that $u_i>0$ in $\bar{U}\times[0,t_0]$ and $z(t)=u_1(t)-u_2(t)$ changes sign in $\tilde{U}$ for $t\in[0,t_0]$. Since $u_{i\varepsilon}\to u_i$ and $z_\varepsilon\to z$ in $C(\bar{\Omega}\times[0,t_0])$ there exist positive numbers μ_0, ν_0 and ε_0 such that for all $\varepsilon\in(0,\varepsilon_0)$

$$u_{i\varepsilon}\geqq\nu_0\quad \text{in } \bar{U}\times[0,t_0],\quad i=1,2, \tag{4.5}$$

and

$$\left\|z_\varepsilon^{\pm}(t)\right\|_{L^\infty(\tilde{U})}\geqq 2\mu_0\quad \text{for } t\in[0,t_0]. \tag{4.6}$$

By (4.5) and the definition of a_ε, condition (d) is satisfied on $\bar{U}\times[0,t_0]$.

Let $\delta\in(0,t_0)$ be fixed.

When $N=1$, Lemma 7.7 below, combined with (4.5), implies that $u_{i\varepsilon}$ is uniformly bounded in $L^\infty(\delta,t_0;H^1(U))$. Hence it follows from the definition of a_ε that condition (b) is satisfied for all $t_1\in(\delta,t_0]$ (the hypothesis H3 is necessary in the proof of Lemma 7.7).

When $N \geqq 2$ we may assume that v_ε is uniformly bounded in $W^{2,p}(\Omega)$. It follows from (4.5) and [16, Thm. 3.1, p. 437] that $u_{i\varepsilon}$ is uniformly bounded in $L^\infty(\delta, t_0, W^{1,\infty}(U))$. Thus condition (c) is satisfied for all $t_1 \in (\delta, t_0]$.

It remains to show that for some $t_1 \in (\delta, t_0]$ condition (a) is satisfied. In view of the conditions (b) and (c), which we proved to be satisfied for $t_1 \in (\delta, t_0]$ we deduce from [16, Thm. 7.1, p. 181] that $\tilde{T}_\varepsilon^\delta(t) z_\varepsilon^\pm(\delta)$ is uniformly bounded in $\tilde{U} \times [\delta, t_0]$. In addition, (4.5) and [16, Thm. 1.1, p. 419] imply that $u_{i\varepsilon}(\delta)$ is uniformly Hölder continuous in $\tilde{U}$. Finally it follows from [16, Thm. 10.1., p. 204] that $T_\varepsilon^\delta(t) z_\varepsilon^\pm(\delta)$ is Hölder continuous in $\tilde{U} \times [\delta, t_0]$, uniformly with respect to $\varepsilon \in (0, \varepsilon_0)$. Hence, by (4.6), there exists a $t_1 \in (\delta, t_0]$ such that condition (a) of Lemma 4.3 is satisfied for all $\varepsilon \in (0, \varepsilon_0)$.

This completes the proof of Theorem 4.1.

5. Stabilization to equilibrium. In the present section we prove the main result of this paper, namely that u stabilizes to equilibrium as $t \to \infty$.

Let the set E be defined by

$$E = \left\{ q \in C(\bar{\Omega}) \colon q \geqq 0 \text{ and } \int_\Omega (\varphi(q)\Delta\eta - q \operatorname{grad} v \operatorname{grad} \eta) = 0 \right.$$
$$\left. \text{for all } \eta \in C^2(\bar{\Omega}) \text{ with } \frac{\partial \eta}{\partial \nu} = 0 \text{ on } \partial\Omega \right\}.$$

It follows from the definition of a solution of Problem P (Definition 3.1) that

$$E = \{ q \in C(\bar{\Omega}) \colon q \geqq 0 \text{ and } u(t; q) = q \text{ for } t \geqq 0 \}. \tag{5.1}$$

Let S be defined by (1.1).

THEOREM 5.1. *If the hypotheses* H1, H2abc, H3 *and* H4 *are satisfied, then*:

(i) $E = S$.

(ii) *There exists a function* $q \in E$ *such that*

$$u(t; u_0) \to q \quad \text{in } C(\bar{\Omega}) \text{ as } t \to \infty$$

where q *satisfies*

$$\int_\Omega q\,dx = \int_\Omega u_0\,dx. \tag{5.2}$$

Remark 5.2. For some functions v and initial functions u_0, condition (5.2) characterizes q completely (see [4]).

The main tools in the proof of Theorem 5.1 are the contraction property which we proved in §4, and the following lemma about the structure of the set S.

LEMMA 5.3. *Let* $q \in C(\bar{\Omega})$ *be nonnegative. Then either* $q \in S$, *or there exists a function* $w \in S$ *such that* $w - q$ *changes sign in a connected subdomain* $U \subset \Omega$ *such that* w, $q > 0$ *in* $\bar{U}$.

Thus S is a continuum in the space of nonnegative continuous functions on $\bar{\Omega}$.

Proof of Lemma 5.3. Suppose that there is no $w \in S$ such that $w - q$ changes sign in some connected subdomain $U \subset \Omega$ such that w, $q > 0$ in $\bar{U}$. We shall prove that $q \in S$.

If $q \equiv 0$ in Ω, then $q \in S$. So let $q(x_1) > 0$ for some $x_1 \in \Omega$. We set $C_1 = \Phi(q(x_1)) + v(x_1)$, where the function Φ is defined by (1.2). Let $P_1 \subset \bar{\Omega}$ be the connected component of the set $\{ x \in \bar{\Omega} \colon v(x) < C_1 \}$ which contains x_1. We claim that

$$\Phi(q(x)) = C_1 - v(x) \quad \text{in } P_1. \tag{5.3}$$

Suppose that P_1 contains a point where $\Phi(q) < C_1 - v$. Then

$$\Phi(q(\tilde{x})) + v(\tilde{x}) = C_1 - \varepsilon_0 \quad \text{and} \quad q(\tilde{x}) > 0$$

for some $\tilde{x} \in P_1$ and $\varepsilon_0 > 0$. Let $\tilde{P}_\varepsilon \subset P_1 (0 < \varepsilon < \varepsilon_0)$ be the connected component of the set $\{x \in \bar{\Omega}: v(x) < C_1 - \varepsilon\}$ which contains $\tilde{x}$. We fix $\varepsilon \in (0, \varepsilon_0)$ so small, that $\tilde{P}_\varepsilon$ contains x_1. Define w by

$$\Phi(w(x)) = \begin{cases} C_1 - \varepsilon - v(x) & \text{for } x \in \tilde{P}_\varepsilon, \\ 0 & \text{for } x \in \bar{\Omega} \setminus \tilde{P}_\varepsilon. \end{cases}$$

Then $w \in S$. Let Γ be a curve in $\tilde{P}_\varepsilon$ which connects $\tilde{x}$ and x_1. Since $w > 0$ on Γ, and since, by construction, $w - q$ changes sign on Γ, there exists a connected closed subset $\Gamma_0 \subset \Gamma$ such that

$$w, q > 0 \text{ on } \Gamma_0 \text{ and } w - q \text{ changes sign on } \Gamma_0.$$

Hence there exists a neighbourhood $\bar{U}$ of Γ_0 in $\tilde{P}_\varepsilon$, where $w, q > 0$ and $w - q$ changes sign. Thus we have a contradiction and P_1 does not contain points where $\Phi(q) < C_1 - v$.

A similar, but easier proof yields that P_1 does not contain points where $\Phi(q) > C_1 - v$, and (5.3) follows.

If $\bar{P}_1 = \bar{\Omega}$ or if $q \equiv 0$ in $\bar{\Omega} \setminus P_1$, then $q \in S$. So suppose that $q(x_2) > 0$ in $\Omega \setminus P_1$. Set $C_2 = \Phi(q(x_2)) + v(x_2)$ and let $P_2 \subset \bar{\Omega}$ be the connected component of the set $\{x \in \bar{\Omega}: C_2 - v(x) > 0\}$ which contains x_2. Then again we conclude that

$$\Phi(q(x)) = C_2 - v(x) \quad \text{in } P_2$$

and clearly $P_1 \cap P_2 = \varnothing$.

Continuing this process, we construct sets, P_i, $i = 1, 2, \cdots$. Since v has a local strict minimum in each connected P_i and since the number of local strict minima of v in Ω is finite, this process is finite. Thus $q \in S$.

Proof of Theorem 5.1(i). We first show that $S \subset E$. Let $w \in S$. Since v has a finite number of local strict mimima, it follows from (1.1) that there exist a finite number of continuous functions $\Phi_i(x)$ $(i = 1, \cdots, i_0)$ with connected and mutually disjoint support such that

$$\Phi(w(x)) = \sum_{i=1}^{i_0} \Phi_i(x) \tag{5.4}$$

and

$$\Phi_i(x) = C_i - v(x) \quad \text{for } x \in \operatorname{supp} \Phi_i$$

for some constants C_i. Since $v \in W^{1,\infty}(\Omega)$ it follows (see for instance [14, Thm. A1, p. 50]) that $\Phi(w(\cdot)) \in W^{1,\infty}(\Omega)$ and

$$\operatorname{grad} \Phi(w) = \begin{cases} -\operatorname{grad} v & \text{in } \{x: w(x) > 0\}, \\ 0 & \text{elsewhere.} \end{cases}$$

Then, by standard theory, $\varphi(w(\cdot)) \in W^{1,\infty}(\Omega)$ and

$$\operatorname{grad} \varphi(w) = \begin{cases} -w \operatorname{grad} v & \text{in } \{x: w(x) > 0\}, \\ 0 & \text{elsewhere.} \end{cases} \tag{5.5}$$

Let $\eta \in C^2(\bar{\Omega})$ with $\partial\eta/\partial\nu = 0$ on $\partial\Omega$. Then, by (5.5),

$$\int_\Omega (\varphi(w)\Delta\eta - w \operatorname{grad} v \operatorname{grad}\eta) = -\int_\Omega (\operatorname{grad}\varphi(w) + w \operatorname{grad} v)\operatorname{grad}\eta = 0.$$

Thus $w \in E$.

Next we show that $E \subset S$. Let $q \in E$ and suppose that $q \notin S$. Then, by Lemma 5.3, there exists a $w \in S$ such that $w - q$ changes sign in a connected subdomain $U \subset \Omega$ in which w, $q > 0$. Since $q \in E$ and $w \in S \subset E$, it follows from (5.1) and Theorem 4.1(ii) that

$$\|q - w\|_{L^1(\Omega)} = \|u(t;q) - u(t;w)\|_{L^1(\Omega)} < \|q - w\|_{L^1(\Omega)}, \qquad t > 0.$$

Thus we have obtained a contradiction and $q \in S$.

Remark 5.4. When $N = 1$, Theorem 5.1(i) follows at once by integrating the differential equation (see [4]).

Proof of Theorem 5.1(ii). We define the ω-limit set

$$\omega(u_0) = \{ w \in L^1(\Omega) : \text{there exists a sequence } t_n \to \infty \text{ such that } u(t_n) \to w \text{ in } L^1(\Omega) \text{ as } t_n \to \infty \}.$$

By Proposition 3.3, the set $\{u(t;u_0); t \geqq 1\}$ is precompact in $C(\bar{\Omega})$ (and hence in $L^1(\Omega)$). Thus $\omega(u_0)$ is nonempty and $\omega(u_0) \subset C(\bar{\Omega})$.

Let $q \in \omega(u_0)$. We show first that q satisfies (5.2), then that $q \in E$, and finally that $\omega(u_0) = \{q\}$.

Setting $\psi(x,t) \equiv 1$ in Definition 3.1, we find that

$$\int_\Omega u(t;u_0) = \int_\Omega u_0 \quad \text{for all } t > 0$$

and (5.2) follows.

In order to show that $q \in E$, we argue by contradiction: suppose that $q \notin E$. Then, by Theorem 5.1(i), $q \notin S$. Thus, by Lemma 5.3, there exists a function $w \in S$ such that $q - w$ changes sign in a connected subdomain $U \subset \Omega$ in which q, $w > 0$. We use w to define the functional V: $L^1(\Omega) \to [0, \infty)$:

$$V(u) = \|u - w\|_{L^1(\Omega)}, \qquad u \in L^1(\Omega).$$

Since $w \in E$, it follows from Theorem 4.1(i) that the solution $u(t)$ of Problem P satisfies

$$V(u(t_1)) \leqq V(u(t_2)) \quad \text{for all } t_1 \geqq t_2 \geqq 0.$$

Thus V is a Lyapunov functional for Problem P. Since $u \in C([0,\infty);\ L^1(\Omega))$ and V is continuous, it follows from [5, Prop. 2.1 and 2.2] that $u(t;q) \in \omega(u_0)$ and that V is constant on $\omega(u_0)$. Hence

$$V(u(t;q)) = V(q) \quad \text{for all } t \geqq 0. \tag{5.6}$$

On the other hand, since q and $w \in C(\bar{\Omega})$, it follows from the choice of w and Theorem 4.1(ii) that

$$V(u(t;q)) < V(q) \quad \text{for all } t > 0$$

which contradicts (5.6). Thus $q \in E$.

Finally we show that $\omega(u_0) \equiv \{q\}$.

Suppose that $\tilde{q}\in\omega(u_0)$ and that $u(t_n;u_0)\to q$ as $t_n\to\infty$ and $u(s_n;u_0)\to\tilde{q}$ as $s_n\to\infty$ where the sequences $\{t_n\}$ and $\{s_n\}$ are chosen such that $s_n<t_n$ for all $n\geqq 1$. Then, using Theorem 4.1(i), we find that

$$\|q-\tilde{q}\|_{L^1(\Omega)}=\lim_{n\to\infty}\|u(t_n;u_0)-\tilde{q}\|_{L^1(\Omega)}$$
$$\leqq\lim_{n\to\infty}\|u(s_n;u_0)-\tilde{q}\|_{L^1(\Omega)}=0.$$

Thus $\tilde{q}=q$, which completes the proof of Theorem 5.1.

6. The Dirichlet problem. In this section we show how the results about Problem P can be extended to the case of homogeneous Dirichlet boundary conditions. We consider the problem

$$(\mathrm{P}_D)\qquad\begin{aligned} &u_t=\Delta\varphi(u)+\operatorname{div}(u\operatorname{grad}v) && \text{in } \Omega\times\mathbb{R}^+,\\ &u=0 && \text{on } \partial\Omega\times\mathbb{R}^+,\\ &u(x,0)=u_0(x) && \text{in } \Omega.\end{aligned}$$

We define a (generalized) solution $u(t;u_0)$ of Problem P_D in a similar way as for Problem P, taking test functions $\psi\in C^{2,1}(\bar{Q})$ such that $\psi=0$ on $\partial\Omega\times\mathbb{R}^+$. The Propositions 3.2, 3.3 and 3.4 as well as Theorem 4.1 remain valid in the case of Problem P_D. In particular

$$u(t;u_0)\in C(\bar{\Omega}) \text{ and } u(t;u_0)=0 \text{ on } \partial\Omega \text{ for } t>0. \tag{6.1}$$

Let the set of steady-state solutions, E_D, be defined by

$$E_D=\{q\in C(\bar{\Omega}):\ q\geqq 0 \text{ and } \textstyle\int_\Omega(\varphi(q)\Delta\eta-q\operatorname{grad}v\operatorname{grad}\eta)=0$$
$$\text{for all } \eta\in C^2(\bar{\Omega}) \text{ such that } \eta=0 \text{ on } \partial\Omega\}.$$

LEMMA 6.1. *Let E be defined as in* §5. *Then $E_D\subset E$.*

Proof. Let $q\in E_D$. Then

$$u(t;q)=q \quad \text{for } t>0. \tag{6.2}$$

Let $\tilde{u}(t;q)$ denote the solution of Problem P with initial function q. Since, by (6.1) and (6.2),

$$\tilde{u}(t;q)\geqq u(t;q)=q=0 \quad \text{on } \partial\Omega\times\mathbb{R}^+,$$

$\tilde{u}(t;q)$ is a supersolution of Problem P_D. Hence, by (6.2),

$$\tilde{u}(t;q)\geqq q \quad \textit{in } \Omega\times\mathbb{R}^+. \tag{6.3}$$

On the other hand we have that

$$\int_\Omega\tilde{u}(t;q)\,dx=\int_\Omega q\,dx, \qquad t\geqq 0.$$

Combined with (6.3) this yields $\tilde{u}(t;q)=q$ for $t\geqq 0$. Thus $q\in E$. We define S_D by

$$S_D=\{q\in S, \text{ such that } q=0 \text{ on } \partial\Omega\},$$

where S is defined by (1.1). In what follows we prove the following theorem.

THEOREM 6.2. *Let the hypotheses* H1, H2abc, H3 *and* H4 *be satisfied. Then*:
(i) $E_D = S_D$.
(ii) E_D *contains a maximal element* $q_{\max}$, *i.e.* $q \leqq q_{\max}$ *in* Ω *for any* $q \in E_D$.
(iii) *There exists a function* $q \in E_D$ *such that* $u(t; u_0) \to q$ *in* $C(\bar{\Omega})$ *as* $t \to \infty$.
If in addition $u_0 \leqq q_{\max}$, *then* q *satisfies* $\int_\Omega q\,dx = \int_\Omega u_0\,dx$.

Proof. (i) By Lemma 6.1 and Theorem 5.1(i), $E_D \subset S$. Hence $E_D \subset S_D$. The proof of the inclusion $S_D \subset E_D$ is identical to the proof of $S \subset E$, given in §5. Thus $E_D = S_D$.

(ii) Let $C > \|v\|_{L^\infty(\Omega)}$ be constant and let $w \in S$ be defined by $\Phi(w(x)) = C - v(x)$, $x \in \bar{\Omega}$. Then $w > 0$ in $\bar{\Omega}$ and it follows from the definition of the set S_D that $q \leqq w$ in Ω for any $q \in S_D$. Hence, by (i),

$$q \leqq w \quad \text{in } \Omega \text{ for any } q \in E_D. \tag{6.4}$$

Since $w \in S = E$, w is a supersolution of Problem P_D. Hence the solution $u(t; w)$ of Problem P_D is nonincreasing in t and we may define

$$0 \leqq p(x) = \lim_{t \to \infty} u(x, t; w), \qquad x \in \bar{\Omega}.$$

By (6.4) and the comparison principle

$$q \leqq p \quad \text{in } \Omega \text{ for any } q \in E_D. \tag{6.5}$$

Below we prove that $p \in E_D$. Then the result follows at once from (6.5), with $q_{\max} = p$.

Let $\eta(x) \geqq 0$ be a smooth test function on $\bar{\Omega}$ such that $\eta = 0$ on $\partial\Omega$. Then $u = u(\cdot; w)$ satisfies

$$\int_\Omega u(t)\eta = \int_\Omega w\eta + \iint_{Q_t} (\varphi(u)\Delta\eta - u \operatorname{grad} v \operatorname{grad} \eta).$$

Thus

$$\frac{d}{dt}\int_\Omega u(t)\eta = \int_\Omega (\varphi(u(t))\Delta\eta - u(t)\operatorname{grad} v \operatorname{grad} \eta). \tag{6.6}$$

Since $u(t; w)$ decreases to p as $t \to \infty$, the left-hand side of (6.6) is nonpositive and there exists a sequence $t_n \to \infty$ such that

$$\frac{d}{dt}\int_\Omega u(t_n)\eta \to 0 \quad \text{as } t_n \to \infty. \tag{6.7}$$

On the other hand, the right-hand side of (6.6) converges to

$$\int_\Omega (\varphi(p)\Delta\eta - p \operatorname{grad} v \operatorname{grad} \eta) \quad \text{as } t_n \to \infty.$$

Hence, by (6.6) and (6.7),

$$\int_\Omega (\varphi(p)\Delta\eta - p \operatorname{grad} v \operatorname{grad} \eta) = 0$$

and thus $p \in E_D$.

(iii) Given an initial function u_0, one can find a function $w \in S$ such that $u_0 \leqq w$ in Ω. Using the above argument and the comparison principle, we find that

$$\limsup_{t \to \infty} u(x, t; u_0) \leqq q_{\max}(x), \qquad x \in \Omega.$$

Hence

$$q \in \omega(u_0) \text{ implies that } q \leqq q_{\max}. \tag{6.8}$$

In order to prove that $u(t;u_0)$ stabilizes to equilibrium, we use the same arguments as for the proof of Theorem 5.1(ii), but now based on the fact that S_D is a continuum between zero and $q_{\max}$, on (6.8), and on the contraction property of u. If furthermore $u_0 \leqq q_{\max}$, then the solution $\tilde{u}(t;u_0)$ of Problem P satisfies $\tilde{u}(t;u_0) \leqq q_{\max}$ for $t \geqq 0$ and in particular $\tilde{u}(t,u_0)=0$ on $\partial\Omega \times \mathbb{R}^+$. Thus $\tilde{u}(t;u_0)$ coincides with the solution $u(t;u_0)$ of Problem P_D (see Lemma 9.3 below). Then, if $q=\lim_{t\to\infty} u(t;u_0)$, we have, by Theorem 5.1(ii)

$$\int_\Omega q\,dx = \int_\Omega u_0\,dx.$$

Part II

7. Existence and regularity. In this section we prove the existence of a solution of Problem P which satisfies this problem in a somewhat stronger sense than that of Definition 3.1. We first recall some usual definitions and then give an alternative definition of a solution, involving the gradient of $\varphi(u)$. The existence proof itself is based on the study of uniformly parabolic problems which are related to Problem P.

We denote by $L^2(0,T;H^1(\Omega))$ the Hilbert space with inner product

$$(u,v)_{L^2(0,T;H^1(\Omega))} = \iint_{Q_T} uv + \iint_{Q_T} \operatorname{grad} u \operatorname{grad} v$$

and by $V_2(Q_T)$ the Banach space with norm

$$|u|^2_{V_2(Q_T)} = \operatorname*{ess\,sup}_{0\leqq t\leqq T} \int_\Omega u^2(t) + \iint_{Q_T} (\operatorname{grad} u)^2.$$

DEFINITION 7.1. We say that $u\colon [0,\infty) \to L^1(\Omega)$ is a *weak solution* of Problem P if it satisfies

(i) $u \in C([0,t];L^1(\Omega)) \cap L^\infty(Q_t)$ for all $t \in (0,\infty)$;

(ii) $\varphi(u) \in V_2(Q_t)$ for all $t \in (0,\infty)$;

(iii) $\int_\Omega u(t)\psi(t) = \int_\Omega u_0\psi(0) + \int_0^t \int_\Omega \{u\psi_t - (\operatorname{grad}\varphi(u) + u \operatorname{grad} v)\operatorname{grad}\psi\}$ for all $\psi \in C^1(\overline{Q})$ and all $t \in (0,\infty)$.

LEMMA 7.2. *A weak solution of Problem* P *is a generalized solution as well.*

Proof. Take $\psi \in C^{2,1}(\overline{Q})$ with $\partial\psi/\partial\nu = 0$ on $\partial\Omega \times \mathbb{R}^+$ and integrate by parts.

In what follows, we show that Problem P has a weak solution. To that purpose, we consider the problems

$$\begin{aligned} &u_t = \Delta\varphi_\varepsilon(u) + \operatorname{div}(u \operatorname{grad} v_\varepsilon) && \text{in } Q_T, \\ &\frac{\partial}{\partial\nu}\varphi_\varepsilon(u) + u\frac{\partial v_\varepsilon}{\partial\nu} = 0 && \text{on } \partial\Omega \times (0,T], \\ &u(x,0) = u_{0\varepsilon}(x) && \text{in } \Omega, \end{aligned}$$

where

$$\varphi_\varepsilon \in C^\infty(\mathbb{R}^+), \qquad \varphi_\varepsilon(0)=0,\ \varphi_\varepsilon'(s) \geqq C(\varepsilon) > 0 \quad \text{for } s \in [0,K],$$
$$(\varphi_\varepsilon^{-1}(s))' \leqq (\varphi^{-1}(s))' \text{ for } s \in [0,\varphi(2K)]$$

where K is the uniform L^∞-bound of u_ε that we find in Lemma 7.4 below and φ_ε and φ_ε' converge to φ and φ' on all compact subsets of $\mathbb{R}^+$ as $\varepsilon \downarrow 0$, where

$$v_\varepsilon \in C^\infty(\bar{\Omega}), \quad \|v_\varepsilon\|_{C^1(\bar{\Omega})} \leqq C \quad \text{for some constant } C > 0,$$
$$\|v_\varepsilon\|_{C(\bar{\Omega})} \leqq \|v\|_{L^\infty(\Omega)} \text{ and } \|v_\varepsilon - v\|_{H^1(\Omega)} \to 0 \quad \text{as } \varepsilon \downarrow 0,$$

and where

$$u_{0\varepsilon} \in C^\infty(\bar{\Omega}), \ 0 \leqq u_{0\varepsilon} \leqq \|u_0\|_{L^\infty(\Omega)},$$

$u_{0\varepsilon}$ satisfies the compatibility condition

$$\frac{\partial}{\partial \nu} \varphi_\varepsilon(u_{0\varepsilon}) + u_{0\varepsilon} \frac{\partial v_\varepsilon}{\partial \nu} = 0 \text{ on } \partial\Omega$$

and $\|u_{0\varepsilon} - u_0\|_{L^2(\Omega)} \to 0$ as $\varepsilon \downarrow 0$.

Since it is standard that one can construct the approximations φ_ε of the function φ having the properties indicated above, we do not do it here. On the other hand we construct explicitly in an appendix approximations v_ε and $u_{0\varepsilon}$ of the functions v and u_0.

To begin with, we give a comparison principle, which turns out to be basic in the study of Problems P_ε and P.

LEMMA 7.3. *Let u_1 and $u_2 \in C^{2,1}(\bar{Q}_T)$ be two solutions of Problem P_ε with initial functions $u_{01} \leqq u_{02}$. Then $u_1(t) \leqq u_2(t)$.*

Proof. Let $z = u_1 - u_2$. Then z satisfies the linear problem L_ε which we discussed in §4 and Lemma 7.3 follows from the comparison principle for that problem.

Before proving the existence of a solution of Problem P, we first give some a priori estimates.

LEMMA 7.4. *Let $u_\varepsilon \in C^{2,1}(\bar{Q}_T)$ be a solution of Problem P_ε. Then*

$$0 \leqq u_\varepsilon \leqq K \quad \text{in } \bar{Q}_T \tag{7.1}$$

where the constant K does not depend on T.

The lower bound in (7.1) is obvious. The upper bound follows from the construction of time-dependent supersolutions, which are uniformly bounded with respect to ε. We leave the details to the reader.

LEMMA 7.5. *Problem P_ε has a unique classical solution $u_\varepsilon \in C^{2+\alpha}(\bar{Q}_T)$ for each $\alpha \in (0,1)$.*

Proof. See [16, Thm. 7.4, p. 491].

In what follows we give some more a priori estimates for u_ε.

LEMMA 7.6. *Let $0 \leqq t - \tau < t < T$. Then there exists $C(\tau) > 0$ such that*

$$\int_{t-\tau}^{t} \int_\Omega (\operatorname{grad} \varphi_\varepsilon(u_\varepsilon))^2 \leqq C(\tau).$$

In particular the constant $C(\tau)$ does not depend on T or on ε.

The proof is immediate if we multiply the differential equation by $\varphi_\varepsilon(u_\varepsilon)$ and integrate by parts over $\Omega \times (t, t+\tau)$.

Next we give an estimate which is useful for the proof of Theorem 4.1(ii); we adapt a proof from Gagneux [10].

LEMMA 7.7. *We suppose that either* $\varphi(s)=s|s|/2$ *or* $\Delta v \in L^1(\Omega)$. *Then*

$$\|\varphi_\varepsilon(u_\varepsilon)\|_{L^\infty(T-\tau,T;H^1(\Omega))} \leqq C(\tau), \qquad 0<\tau\leqq T. \tag{7.2}$$

The constant $C(\tau)$ *does not depend on* T.

Proof. We first show that for $0\leqq t-\tau<t\leqq T$, the following estimate holds

$$\int_{t-\tau}^{t}\int_\Omega \varphi_\varepsilon'(u_\varepsilon)(\operatorname{grad} u_\varepsilon)^2 \leqq C(\tau). \tag{7.3}$$

For that purpose we multiply the differential equation by u_ε and integrate by parts; we obtain

$$\frac{1}{2}\int_\Omega u_\varepsilon^2(t)-\frac{1}{2}\int_\Omega u_\varepsilon^2(t-\tau)+\int_{t-\tau}^{t}\int_\Omega \varphi_\varepsilon'(u_\varepsilon)(\operatorname{grad} u_\varepsilon)^2=\int_{t-\tau}^{t}\int_\Omega \operatorname{grad} v_\varepsilon \operatorname{grad}\left(\frac{1}{2}u_\varepsilon^2\right).$$

When (i): $\varphi(s)=s^2/2$, we have

$$\int_{t-\tau}^{t}\int_\Omega \operatorname{grad} v_\varepsilon \operatorname{grad}\left(\frac{1}{2}u_\varepsilon^2\right) \leqq C\sqrt{\tau}\,\|\varphi(u_\varepsilon)\|_{L^2(t-\tau,t;H^1(\Omega))} \leqq C(\tau)$$

by Lemma 7.4; and when (ii): $\Delta v\in L^1(\Omega)$, then

$$\int_{t-\tau}^{t}\int_\Omega \operatorname{grad} v_\varepsilon \operatorname{grad}\left(\frac{1}{2}u_\varepsilon^2\right)=\int_{t-\tau}^{t}\int_{\partial\Omega}\frac{u_\varepsilon^2}{2}\frac{\partial v_\varepsilon}{\partial \nu}-\int_{t-\tau}^{t}\int_\Omega \Delta v_\varepsilon \frac{u_\varepsilon^2}{2}\leqq C(\tau)$$

which completes the proof of (7.3).

For the rest of the proof we refer to [10].

We shall need a result of DiBenedetto [7, Thm. 6.2] to deduce a strong estimate, namely the equicontinuity of u_ε.

LEMMA 7.8(i). *For every* $\tau>0$ *there exists a continuous nondecreasing function* $\omega_\tau(\cdot)$, $\omega_\tau(0)=0$ *such that*

$$|u_\varepsilon(x_1,t_1)-u_\varepsilon(x_2,t_2)|\leqq \omega_\tau\left(|x_1-x_2|+|t_1-t_2|^{1/2}\right)$$

for all $(x_i,t_i)\in\bar\Omega T\times[\tau,T]$, $i=1,2$. *The function* ω_τ *does not depend on* T *and* ε.

(ii) *If* $u_0\in C(\bar\Omega)$, *then* $\{u_\varepsilon\}$ *is equicontinuous on* $\bar\Omega\times[0,T]$.

We are now in a position to prove the existence theorem.

THEOREM 7.9. *We suppose that* H1 *and* H4 *are satisfied and that* $v\in W^{1,\infty}(\Omega)$. *Then there exists a weak solution of Problem* P *which satisfies*

$$0\leqq u\leqq C \quad \text{on } Q_T$$

and is continuous in any set $\bar\Omega\times[\tau,T]$ *with* $\tau>0$. *The constant* C *and the modulus of continuity do not depend on* T.

Proof. From the estimates above we deduce that there exist a function $u\in L^\infty(Q_T)\cap C(\bar\Omega\times(0,T])$ and a subsequence of $\{u_\varepsilon\}$ which we denote again by $\{u_\varepsilon\}$ such that

(i) $u_\varepsilon\to u$ uniformly on all sets of the form $\bar\Omega\times[\tau,T]$ with $\tau>0$ (by Lemma 7.8),

(ii) $u_\varepsilon\to u$ strongly in $L^2(Q_T)$ and a.e. (this is a consequence of (i) and the uniform bound of u_ε in $L^\infty(\Omega)$),

(iii) $\varphi_\varepsilon(u_\varepsilon)\to\varphi(u)$ weakly in $L^2(0,T;H^1(\Omega))$ (this follows from Lemma 7.6; one checks that the limit is $\varphi(u)$ by observing that by (ii) and Lebesgue's dominated convergence theorem $\varphi_\varepsilon(u_\varepsilon)\to\varphi(u)$ strongly in $L^2(Q_T)$),

(iv) $u_\varepsilon \operatorname{grad} v_\varepsilon \to u \operatorname{grad} v$ strongly in $L^1(Q_T)$.

It remains to check that u is a solution of Problem P. It is easy to deduce from (i)–(iv) that u satisfies the integral equation in Definition 3.1 since u_ε satisfies a similar equation. Also $u \in C((0,T];\ L^1(\Omega))$. In order to show that $\|u(t)\|_{L^1(\Omega)}$ is continuous at zero we use the contraction Theorem 4.1(i). Let $\tilde{u}_\varepsilon$ be a solution of Problem P with initial function $u_{0\varepsilon}$ obtained as a limit of solutions of Problem P_ε. Then

$$\|u(t)-u_0\|_{L^1(\Omega)} \leqq \|u(t)-\tilde{u}_\varepsilon(t)\|_{L^1(\Omega)} + \|\tilde{u}_\varepsilon(t)-u_{0\varepsilon}\|_{L^1(\Omega)} + \|u_{0\varepsilon}-u_0\|_{L^1(\Omega)}.$$

Let $\eta > 0$ be arbitrary. Since $u_{0\varepsilon}$ converges to u_0 in $L^1(\Omega)$, one can fix ε such that $\|u_{0\varepsilon}-u_0\|_{L^1(\Omega)} \leqq \eta/3$. Then by Theorem 4.1 $\|u(t)-\tilde{u}_\varepsilon(t)\|_{L^1(\Omega)} \leqq \eta/3$. Finally we deduce from Lemma 7.8 (ii) that one can find t_0 such that $\|\tilde{u}_\varepsilon(t)-u_{0\varepsilon}\|_{L^1(\Omega)} \leqq \eta/3$ for all $t \leqq t_0$.

Remark 7.10. If the function φ is defined on $\mathbb{R}$ with $\varphi'(s) > 0$ for $s < 0$, the condition $u_0 \geqq 0$ is not necessary to obtain the results of §7.

8. Uniqueness of the solution. In order to show that the solution of Problem P is unique, we apply a method due to Kalashnikov [13] which consists of comparing an arbitrary solution of Problem P with a solution obtained as the limit of a sequence of classical solutions of the parabolic equation in Problem P. We do so below and for technical reasons which will appear later we impose the condition $\Delta v \geqq -M$ in the sense of distributions.

We approximate Problem P in two steps, first by the problem

$$(P_n)\qquad \begin{aligned} &u_t = \Delta\varphi(u) + \operatorname{div}(u \operatorname{grad} v) && \text{in } Q_T,\\ &\frac{\partial}{\partial \nu}\varphi(u) + u\frac{\partial v}{\partial \nu} = \frac{A}{n}e^{-Mt} && \text{on } \partial\Omega \times (0,T],\\ &u(x,0) = u_{0n}(x) := u_0(x) + \frac{1}{n} && \text{in } \Omega, \end{aligned}$$

which in turn we approximate by the problem

$$(P_{nj})\qquad \begin{aligned} &u_t = \Delta\varphi(u) + \operatorname{div}(u \operatorname{grad} v_j) && \text{in } Q_T,\\ &\frac{\partial}{\partial \nu}\varphi(u) + u\frac{\partial v_j}{\partial \nu} = \frac{A}{n}e^{-Mt} && \text{on } \partial\Omega \times (0,T],\\ &u(x,0) = u_{0j}(x) + \frac{1}{n} && \text{in } \Omega, \end{aligned}$$

where

$$\begin{aligned} &v_j \in C^\infty(\bar{\Omega}), \qquad \|v_j\|_{C^1(\bar{\Omega})} \leqq C_1 \quad \text{for some constant } C_1 > 0,\\ &\Delta v_j \geqq -M \text{ and } \|v_j - v\|_{H^1(\Omega)} \to 0 \quad \text{as } j \to \infty, \end{aligned}$$

where the constant A is such that $A \geqq C_1$ and

$$u_{0j} \in C^{2+\alpha}(\bar{\Omega}), \qquad 0 \leqq u_{0j} \leqq C_2 \quad \text{for some constant } C_2 > 0,$$

u_{0j} satisfies the compatibility condition

$$\varphi'\left(u_{0j} + \frac{1}{n}\right)\frac{\partial u_{0j}}{\partial \nu} + \frac{\partial v_j}{\partial \nu}u_{0j} + \frac{1}{n}\left(\frac{\partial v_j}{\partial \nu} - A\right) = 0 \quad \text{on } \partial\Omega$$

and is such that $\|u_{0j}-u_0\|_{L^2(\Omega)} \to 0$ as $j \to \infty$.

We show in the appendix that one can construct such functions v_j and u_{0j}.

We first give uniform upper and lower bounds for the solution u_{nj} of Problem P_{nj}; the fact that u_{nj} turns out to be bounded away from zero ensures that Problem P_{nj} is uniformly parabolic.

LEMMA 8.1. *Let $u_{nj} \in C^{2,1}(\overline{Q}_T)$ be a solution of Problem P_{nj}. Then, for n large enough,*

$$\frac{1}{n} e^{-Mt} \leqq u_{nj}(x,t) \leqq C \quad \text{for all } (x,t) \in \overline{Q}_T$$

where the constant C does not depend on time.

The main tool of the proof is the following comparison principle which is an immediate generalization of Lemma 7.3.

LEMMA 8.2. *Let u_1 and $u_2 \in C^{2,1}(\overline{Q}_T)$ and assume that u_1 and u_2 are positive on $\overline{Q}_T$. If*

$$\Delta\varphi(u_1) + \operatorname{div}(u_1 \operatorname{grad} v_j) - u_{1t} \geqq \Delta\varphi(u_2) + \operatorname{div}(u_2 \operatorname{grad} v_j) - u_{2t} \quad \text{in } Q_T,$$

$$\frac{\partial}{\partial \nu}\varphi(u_1) + u_1 \frac{\partial v_j}{\partial \nu} \leqq \frac{\partial}{\partial \nu}\varphi(u_2) + u_2 \frac{\partial v_j}{\partial \nu} \quad \text{on } \partial\Omega \times (0,T]$$

$$u_1(x,0) \leqq u_2(x,0) \quad \text{in } \Omega.$$

Then

$$u_1 \leqq u_2 \quad \text{in } \overline{Q}_T.$$

Proof of Lemma 8.1. One easily checks that $s^-(x,t) := (1/n)e^{-Mt}$ is a subsolution of Problem P_{nj}, and hence, by Lemma 8.2, $u_n \geqq (1/n)e^{-Mt}$.

On the other hand, one can construct a supersolution of the form

$$s^+(x,t) = \Phi^{-1}\big(C - v_j - e^{-Mt}h(x)\big),$$

where h is a smooth function such that $1 \leqq h \leqq 2$, and where the constant C is chosen large enough. Omitting the details here we find that $u_{nj} \leqq s^+ \leqq \Phi^{-1}(C)$.

By the method of §7 one can obtain further a priori estimates for solutions of the Problems P_{nj} and P_n and use them to show that a subsequence $\{u_{nj}\}$ of solutions of Problems P_{nj} converge to a generalized solution of Problem P_n as $j \to \infty$ and then that a subsequence $\{u_{\tilde{n}}\}$ of solutions of Problems $P_{\tilde{n}}$ converge to a solution of Problem P. In addition, following DiBenedetto again, we find that the sequence $\{u_{nj}\}$ is equicontinuous. In particular one can show that there exists a solution u of Problem P and a subsequence of the solutions u_n of Problems P_n (which we denote again by $\{u_n\}$) which converge to u as $n \to \infty$. Below we use this construction to prove the following result.

THEOREM 8.3. *We suppose that the hypotheses* H1, H2a *and* H4 *are satisfied. Let u be the solution of Problem* P *obtained above and let $\underline{u}$ (resp. $\bar{u}$) be a subsolution (resp. supersolution) of P with initial function $\underline{u}_0$ (resp. $\bar{u}_0$). Then for every $t \in (0,T]$ we have that*

$$\int_\Omega (\underline{u}(t) - u(t))^+ \leqq \int_\Omega (\underline{u}_0 - u_0)^+ \tag{8.1}$$

and

$$\int_\Omega (u(t) - \bar{u}(t))^+ \leqq \int_\Omega (u_0 - \bar{u}_0)^+. \tag{8.2}$$

COROLLARY 8.4. *If the hypotheses* H1, H2a *and* H4 *are satisfied, Problem* P *has a unique solution.*

COROLLARY 8.5. *Let* $\underline{u}(t)$ *and* $\bar{u}(t)$ *be respectively a subsolution and a supersolution of Problem* P *with initial functions* $\underline{u}_0$ *and* $\bar{u}_0$ *such that* $\underline{u}_0 \leqq \bar{u}_0$. *Then* $\underline{u}(t) \leqq \bar{u}(t)$ *for every* $t \in (0, T]$.

Proof of Theorem 8.3. The proof follows closely that of Diaz and Kersner [6]. Let ψ be a test function. Then

$$\int_\Omega (\underline{u}-u_n)(t)\psi(t) - \int_\Omega (\underline{u}_0 - u_{0n})\psi(0)$$

$$\leqq \int_0^t \int_\Omega \{(\underline{u}-u_n)(t)\psi_t + (\varphi(\underline{u})-\varphi(u_n))\Delta\psi - (\underline{u}-u_n)\operatorname{grad} v \operatorname{grad}\psi\}$$

$$- \frac{A}{n}\int_0^t \int_{\partial\Omega} e^{-Mt}\psi$$

$$\leqq \int_0^t \int_\Omega (\underline{u}-u_n)\{\psi_t + A_n\Delta\psi - \operatorname{grad} v \operatorname{grad}\psi\}$$

where

$$A_n(x,t) = \int_0^1 \varphi'(\theta\underline{u}(x,t) + (1-\theta)u_n(x,t))\,d\theta.$$

Since $u_n \geqq (1/n)e^{-Mt}$, there exists $\varepsilon(n) > 0$ such that $A_n \geqq \varepsilon(n) > 0$. We now define a sequence of smooth functions $A_{nj} \geqq \varepsilon(n)$ such that

$$A_{nj} \to A_n \quad \text{strongly in } L^2(Q_T) \text{ as } j \to \infty.$$

Let ψ_{nj} be the solution of the problem

$$(\mathrm{L}_{nj}) \qquad \begin{aligned} &\psi_t + A_{nj}\Delta\psi - \operatorname{grad} v_j \operatorname{grad}\psi = 0 && \text{in } \Omega\times(0,t),\\ &\frac{\partial\psi}{\partial\nu} = 0 && \text{on } \partial\Omega\times[0,t),\\ &\psi(x,t) = \chi(x) && \text{in } \Omega, \end{aligned}$$

where χ is a smooth function such that $0 \leqq \chi \leqq 1$. As a consequence of the maximum principle we have that $0 \leqq \psi_{nj} \leqq 1$. We set $\psi = \psi_{nj}$. Then

$$(8.3) \quad \int_\Omega (\underline{u}-u_n)\chi \leqq \int_\Omega (\underline{u}_0 - u_{0n})^+ + \int_0^t \int_\Omega (\underline{u}-u_n)\{(A_n - A_{nj})\Delta\psi_{nj} - (\operatorname{grad} v - \operatorname{grad} v_j)\operatorname{grad}\psi_{nj}\}.$$

In what follows we first keep n fixed. In order to show that the second term of the right-hand side of (8.3) vanishes as $j \to \infty$, it is sufficient to prove that there exists a constant $C(n,t)$ such that

$$(8.4) \qquad \int_0^t \int_\Omega (\operatorname{grad}\psi_{nj})^2 \leqq C(n,t) \quad \text{and} \quad \int_0^t \int_\Omega (\Delta\psi_{nj})^2 \leqq C(n,t).$$

These estimates follow from multiplying the differential equation in Problem L_{nj} by $\Delta\psi_{nj}$ and integrating it on $\Omega\times(0,t)$. For details we refer to Aronson, Crandall and Peletier [3] where a similar calculation is made. Inequality (8.3) together with (8.4)

yields

$$\int_\Omega (\underline{u}(t)-u_n(t))\chi \leqq \int_\Omega (\underline{u}_0-u_{0n})^+$$

for all smooth χ such that $0\leqq\chi\leqq 1$ and hence, since $u_{0n}\to u_0$ in $C(\bar{\Omega})$ and $u_n\to u$ in $C(\bar{Q}_T)$, we have

$$\int_\Omega (\underline{u}(t)-u(t))\chi \leqq \int_\Omega (\underline{u}_0-u_0)^+. \tag{8.5}$$

Next we consider a sequence of smooth functions χ_m such that χ_m converges in $L^2(\Omega)$ to a limit $\bar{\chi}$ defined by

$$\bar{\chi}(x)=\begin{cases}1 & \text{in } \{x|\underline{u}(x,t)\geqq u(x,t)\},\\ 0 & \text{elsewhere.}\end{cases}$$

Taking $\chi=\chi_m$ in (8.5) and letting $m\to\infty$ yield (8.1). Finally one can show (8.2) in a similar way.

9. Some remarks about the Dirichlet problem. In this section we give results about the existence, uniqueness and regularity of solutions of Problem P_D, which we introduced in §6. Because the proofs are quite similar to those given in §8, we omit them here.

THEOREM 9.1 (existence + regularity). *Let* H1 *and* H4 *be satisfied and let* $v\in W^{1,\infty}(\Omega)$. *Then Problem* P_D *possesses a solution* u *which is uniformly bounded in* Q *and which is continuous in any set* $\bar{\Omega}\times[\tau,T]$ *with* $\tau>0$. *The modulus of continuity does not depend on* T.

THEOREM 9.2 (uniqueness + comparison principle). *Let* H1, H2a *and* H4 *be satisfied. Then*:

(i) *Problem* P_D *possesses at most one solution.*

(ii) *Let* $\underline{u}(t)$ *and* $\bar{u}(t)$ *be respectively a subsolution and a supersolution of Problem* P_D *with respect to the initial functions* $\underline{u}_0$ *and* $\bar{u}_0$. *If* $\underline{u}_0\leqq\bar{u}_0$ *in* Ω, *then* $\underline{u}(t)\leqq\bar{u}(t)$ *in* Ω *for* $t\geqq 0$.

Finally we give a lemma which relates solutions of the Neumann problem which vanish on $\partial\Omega$ to solutions of the Dirichlet problem.

LEMMA 9.3. *If the solution* $u(t;u_0)$ *of Problem* P *satisfies* $u(t;u_0)=0$ *on* $\partial\Omega$ *for any* $t\geqq 0$, *then* $u(t;u_0)$ *is a solution of Problem* P_D.

Proof. Let $\psi\in C^{2,1}(\bar{Q})$ with $\psi=0$ on $\partial\Omega\times\mathbb{R}^+$. Then $u(t;u_0)$ satisfies the integral equality (iii) of Definition 7.1. Integrating by parts yields

$$\int_\Omega u(t)\psi(t)=\int_\Omega u_0\psi(0)-\int_0^t\int_{\partial\Omega}\varphi(u)\frac{\partial\psi}{\partial\nu}$$

$$+\int_0^t\int_\Omega\{\varphi(u)\Delta\psi+u\psi_t-u\operatorname{grad}v\operatorname{grad}\psi\}.$$

Since $\varphi(u)=0$ on $\partial\Omega\times\mathbb{R}^+$, the second term at the right-hand side vanishes. Thus $u(t;u_0)$ is a solution of Problem P_D.

Appendix. In this appendix we collect various approximation results which are used in this article.

A1. *Approximation of* v.

LEMMA A1. *Let* $v\in W^{1,\infty}(\Omega)$. *Then there exists a sequence* $\{v_\varepsilon\}\subset C^\infty(\bar{\Omega})$ *such that* $\|v_\varepsilon\|_{C^1(\bar{\Omega})}\leqq C$, $\|v_\varepsilon\|_{C(\bar{\Omega})}\leqq\|v\|_{L^\infty(\Omega)}$ *and* $\|v_\varepsilon-v\|_{H^1(\Omega)}\to 0$ *as* $\varepsilon\downarrow 0$.

Proof. Let $\tilde{\Omega} \supset \Omega$ with $\operatorname{dist}(\Omega, \partial\tilde{\Omega}) > 0$. Then one can extend v by a function $\tilde{v} \in W^{1,\infty}(\tilde{\Omega})$ such that $\tilde{v} = v$ in Ω and $\|\tilde{v}\|_{L^\infty(\tilde{\Omega})} \leqq \|v\|_{L^\infty(\Omega)}$. We define the function

$$\rho(x) = \begin{cases} 0 & \text{if } |x| \geqq 1, \\ C\exp\left\{\dfrac{1}{|x|^2 - 1}\right\} & \text{if } |x| < 1, \end{cases}$$

where C is a constant such that $\int_{\mathbb{R}^n} \rho(x)\,dx = 1$. Let

$$v_\varepsilon(x) = \varepsilon^{-N} \int_{\tilde{\Omega}} \rho\left(\frac{x-y}{\varepsilon}\right) \tilde{v}(y)\,dy \quad \text{for } x \in \tilde{\Omega}.$$

In particular note that $\|v_\varepsilon\|_{C(\bar{\Omega})} \leqq \|v\|_{L^\infty(\Omega)}$. Let us suppose that $\varepsilon < \operatorname{dist}(\Omega, \partial\tilde{\Omega})$. Then it is a standard result (see for instance Kufner et al. [15]) that

$$\operatorname{grad} v_\varepsilon(x) = \varepsilon^{-N} \int_{\tilde{\Omega}} \rho\left(\frac{x-y}{\varepsilon}\right) \operatorname{grad} v(y)\,dy \quad \text{for } x \in \Omega. \tag{A.1}$$

Then $\|\operatorname{grad} v_\varepsilon\|_{C(\bar{\Omega})} \leqq \|\operatorname{grad} v\|_{L^\infty(\tilde{\Omega})}$ and $\|v_\varepsilon - v\|_{W^{1,p}(\Omega)} \to 0$ as $\varepsilon \downarrow 0$ for every $p \in [1, \infty)$.

LEMMA A2. *Let $v \in W^{1,\infty}(\tilde{\Omega})$ be such that $\Delta v \geqq -M$ in the sense of distributions in $\tilde{\Omega} \supset \Omega$ with $\operatorname{dist}(\Omega, \partial\tilde{\Omega}) > 0$. Then there exists a sequence $\{v_\varepsilon\} \subset C^\infty(\bar{\Omega})$ such that $\|v_\varepsilon\|_{C^1(\bar{\Omega})} \leqq C$, $\Delta v_\varepsilon \geqq -M$ in Ω and $\|v_\varepsilon - v\|_{H^1(\Omega)} \to 0$ as $\varepsilon \downarrow 0$.*

Proof. In view of the proof of Lemma A1, it remains to show that $\Delta v_\varepsilon \geqq -M$. From (A.1) we deduce that

$$\Delta v_\varepsilon(x) = \varepsilon^{-N} \left\langle \Delta v(y), \rho\left(\frac{x-y}{\varepsilon}\right) \right\rangle \quad \text{for } x \in \Omega$$

where $\langle \cdot, \cdot \rangle$ denotes the duality pairing between $H_0^1(\tilde{\Omega})$ and $H^{-1}(\tilde{\Omega})$. In particular, since $\Delta v \geqq -M$ we have that for $x \in \Omega$

$$\Delta v_\varepsilon(x) \geqq -\varepsilon^{-N} M \int_{\tilde{\Omega}} \frac{\rho(x-y)}{\varepsilon}\,dy = -\varepsilon^{-N} M \int_{\mathbb{R}} \rho\left(\frac{u}{\varepsilon}\right) du = -M$$

which yields the result.

A2. *Approximation of u_0.*

LEMMA A3. *Let $u_0 \in L^\infty(\Omega)$ with $u_0 \geqq 0$ a.e. and let v_ε, $v_j \in C^\infty(\bar{\Omega})$ be such that $\|v_\varepsilon\|_{C^1(\bar{\Omega})}$, $\|v_j\|_{C^1(\bar{\Omega})} \leqq C$. Then*

(i) *There exists a sequence $\{u_{0\varepsilon}\} \subset C^\infty(\bar{\Omega})$ such that $0 \leqq u_{0\varepsilon} \leqq \|u_0\|_{L^\infty(\Omega)}$, $u_{0\varepsilon}$ satisfies the compatibility condition*

$$\varphi_\varepsilon'(u_{0\varepsilon}) \frac{\partial u_{0\varepsilon}}{\partial \nu} + u_{0\varepsilon} \frac{\partial v_\varepsilon}{\partial \nu} = 0 \text{ on } \partial\Omega \quad \text{and} \quad \|u_{0\varepsilon} - u_0\|_{L^2(\Omega)} \to 0 \text{ as } \varepsilon \downarrow 0.$$

(ii) *There exists a sequence $\{u_{0j}\} \subset C^\infty(\bar{\Omega})$ such that $0 \leqq u_{0j} \leqq C$, u_{0j} satisfies the compatibility condition*

$$\varphi'\left(u_{0j} + \frac{1}{n}\right) \frac{\partial u_{0j}}{\partial \nu} + \frac{\partial v_j}{\partial \nu} u_{0j} + \left(\frac{\partial v_j}{\partial \nu} - A\right) \Big/ n = 0$$

where A is a given constant and $\|u_{0j} - u_0\|_{L^2(\Omega)} \to 0$ as $j \to \infty$.

Proof. Since (i) is practically a special case of (ii) we only prove (ii). We define

$$\tilde{u}_{0j}(x) = j^n \int_\Omega \rho(j(x-y)) u_0(y)\,dy$$

and note that $0 \leqq \tilde{u}_{0j} \leqq \|u_0\|_{L^\infty(\Omega)}$. Let B be a positive constant. It follows from Friedman [9, p. 39] that one can find a function $w_j \in C^\infty(\bar{\Omega})$ such that

$$w_j\Big|_{\partial\Omega} = B \text{ and } \frac{\partial w_j}{\partial \nu}\Big|_{\partial\Omega} = \frac{-1}{\varphi'(B+n^{-1})}\left(\frac{\partial v_j}{\partial \nu}B + \left(\frac{\partial v_j}{\partial \nu} - A\right)/n\right).$$

Also, since $\operatorname{grad} v_j$ is bounded in $C(\bar{\Omega})$ uniformly in j we have that $\|w_j\|_{C(\bar{\Omega})} \leqq C$. Since $B > 0$, there exists $\tilde{\Omega}_j \subset \Omega$ with $\operatorname{dist}(\tilde{\Omega}_j, \partial\Omega) > 0$ such that $w_j > 0$ on $\Omega \setminus \tilde{\Omega}_j$. Finally we choose $\Omega_{1j} \subset \Omega_{2j} \subset \Omega$, such that $\operatorname{dist}(\Omega_{2j}, \partial\Omega) > 0$, $\operatorname{dist}(\Omega_{1j}, \partial\Omega_{2j}) > 0$, $\Omega_{1j} \supset \tilde{\Omega}_j$ and $\operatorname{meas}(\Omega \setminus \Omega_{1j}) \leqq 1/j$. We define

$$u_{0j}(x) = \begin{cases} \tilde{u}_{0j}(x) & \text{if } x \in \Omega_{1j}, \\ \xi_j(x)\tilde{u}_{0j}(x) + (1-\xi_j(x))w_j(x) & \text{if } x \in \Omega_{2j} \setminus \Omega_{1j}, \\ w_j(x) & \text{if } x \in \Omega \setminus \Omega_{2j}, \end{cases}$$

where ξ_j is a C^∞ function such that

$$\xi_j(x) \in \begin{cases} 1 & \text{if } x \in \Omega_{1j}, \\ [0,1] & \text{if } x \in \Omega_{2j} \setminus \Omega_{1j}, \\ 0 & \text{if } x \in \Omega \setminus \Omega_{2j}. \end{cases}$$

We have that $u_{0j} \in C^\infty(\bar{\Omega})$. Also

$$\|u_{0j} - u_0\|_{L^2(\Omega)} \leqq \|u_{0j} - \tilde{u}_{0j}\|_{L^2(\Omega)} + \|\tilde{u}_{0j} - u_0\|_{L^2(\Omega)}.$$

$\|\tilde{u}_{0j} - u_0\|_{L^2(\Omega)}$ can be made arbitrarily small by choosing j large enough. The term $\|u_{0j} - \tilde{u}_{0j}\|_{L^2(\Omega)}$ is bounded by $(1/j)(\|u_0\|_{L^\infty(\Omega)} + \|w_j\|_{L^\infty(\Omega)})$, which tends to zero as $j \to \infty$.

Acknowledgments. The authors are greatly indebted to Professor M. E. Gurtin for suggesting this problem to them and for many discussions. They wish to express their thanks to Professor L. A. Peletier whose advice has been invaluable for the completion of this work.

REFERENCES

[1] N. D. Alikakos and R. Rostamian, *Large time behavior of solutions of Neumann boundary value problem for the porous medium equation*, Indiana Univ. Math. J., 30 (1981), pp. 749–785.

[2] ———, *Stabilization of solutions of the equation* $\partial u/\partial t = \Delta\varphi(u) - \beta(u)$, Nonlinear Anal. TMA, 6 (1982), pp. 637–647.

[3] D. G. Aronson, M. G. Crandall and L. A. Peletier, *Stabilization of solutions of a degenerate nonlinear diffusion problem*, Nonlinear Anal. TMA, 6 (1982), pp. 1001–1022.

[4] M. Bertsch, M. E. Gurtin, D. Hilhorst and L. A. Peletier, *On interacting populations that disperse to avoid crowding: the effect of a sedentary colony*, J. Math. Biol., 19 (1984), pp. 1–12.

[5] C. M. Dafermos, *Asymptotic behavior of solutions of evolution equations*, in Nonlinear Evolution Equations, M. G. Crandall, ed., Academic Press, New York, 1978.

[6] J. I. Diaz and R. Kersner, *On a nonlinear degenerate parabolic equation in infiltration or evaporation through a porous medium*, to appear.

[7] E. DiBenedetto, *Continuity of weak solutions to a general porous media equation*, Indiana Univ. Math. J., 32 (1983), pp. 83–118.

[8] A. Friedman, *Partial Differential Equations of Parabolic Type*, Prentice-Hall, Englewood Cliffs, NJ, 1964.

[9] _____, *Partial Differential Equations*, Holt-Rinehart, New York, 1969.

[10] G. GAGNEUX, *Déplacement de fluides non miscibles incompressibles dans un cylindre poreux*, J. de Mécanique, 19 (1980), pp. 295–325.

[11] M. E. GURTIN AND A. C. PIPKIN, *A note on populations that disperse to avoid crowding*, Quart. Appl. Math., 42 (1984), pp. 87–94.

[12] A. S. KALASHNIKOV, *The propagation of disturbances in problems of nonlinear heat conduction with absorption*, U.S.S.R. Comput. Math. and Math. Phys., 14, 4 (1974), pp. 70–85.

[13] _____, *The Cauchy problem in a class of growing functions for equations of unsteady filtration type*, Vestnik Moskov Univ. Ser. VI Mat. Mech., 6 (1963), pp. 17–27. (In Russian.)

[14] D. KINDERLEHRER AND G. STAMPACCHIA, *An Introduction to Variational Inequalities and Their Applications*, Academic Press, New York, 1980.

[15] A. KUFNER, O. JOHN AND S. FUČIC, *Function Spaces*, Noordhoff, Leiden 1977.

[16] O. A. LADYŽENSKAJA, V. A. SOLONNIKOV AND N. N. URAL'CEVA, *Linear and Quasilinear Equations of Parabolic Type*, Transl. Math. Monographs 23, American Mathematical Society, Providence, RI, 1968.

[17] M. MADAUNE-TORT, *Perturbations singulières de problèmes aux limites du second ordre hyperboliques et paraboliques non linéaires*, Thèse, Université de Pau, 1981.

[18] S. OSHER AND J. RALSTON, *L^1 stability of travelling waves with applications to convective porous media flow*, Comm. Pure Appl. Math., 35 (1982), pp. 737–749.

[19] H. RADEMACHER, *Über partielle und totale Differenzierbarkeid von Funktionen mehrerer Variabeln und über die Transformation der Doppelintegrale*, Math. Ann., 79 (1919), pp. 340–359 (and in Collected Papers of Hans Rademacher, E. Grosswald, ed., MIT Press, Cambridge, MA, 1974).

[20] M. SCHATZMAN, *Stationary solutions and asymptotic behavior of a quasilinear degenerate parabolic equation*, Indiana Univ. Math. J., 33 (1984), pp. 1–29.

[21] H. TOURÉ, *Etude des équations générales $u_t - \varphi(u)_{xx} + f(u)_x = v$ par la théorie des semi-groupes non linéaires dans L^1*, Thèse, Université de Franche-Comté Besançon, 1982.

SIAM J. MATH. ANAL.
Vol. 17, No. 4, July 1986

009

THE WELL-POSEDNESS OF THE KURAMOTO–SIVASHINSKY EQUATION*

EITAN TADMOR†

Abstract. The Kuramoto–Sivashinsky equation arises in a variety of applications, among which are modeling reaction-diffusion systems, flame-propagation and viscous flow problems. It is considered here, as a prototype to the larger class of generalized Burgers equations: those consist of quadratic nonlinearity and arbitrary linear parabolic part. We show that such equations are well-posed, thus admitting a unique smooth solution, continuously dependent on its initial data. As an attractive alternative to standard energy methods, existence and stability are derived in this case, by "patching" in the large short time solutions without "loss of derivatives".

Key words. Kuramoto–Sivashinsky equation, fixed point iterations, existence, uniqueness, stability

AMS(MOS) subject classifications. Primary 35Q20; secondary 35K55

1. Introduction.

The equation referred to in the title is of the form

$$\frac{\partial \phi}{\partial t}+|\nabla \phi|^2+\Delta \phi+\Delta^2 \phi=0.$$

This equation was independently advocated by Kuramoto [2], in connection with reaction-diffusion systems, and by Sivashinsky [4], modeling flame propagation; it also arises in the context of viscous film flow [5] and bifurcating solutions of the Navier–Stokes equations.[1]

In this paper we study the well-posedness question associated with the one-dimensional version of the Kuramoto–Sivashinsky equation (abbreviated hereafter as the K-S equation)

$$\frac{\partial \phi}{\partial t}+\left(\frac{\partial \phi}{\partial x}\right)^2+\frac{\partial^2 \phi}{\partial x^2}+\frac{\partial^4 \phi}{\partial x^4}=0. \tag{1.1}$$

It is shown that the Cauchy problem connected with (1.1) is well-posed: the K-S equation admits a unique smooth solution, continuously dependent on its initial data. In fact, all the results quoted below equally apply to the more general equation

$$\frac{\partial \phi}{\partial t}+\left(\frac{\partial \phi}{\partial x}\right)^2=P\left(\frac{\partial}{\partial x}\right)\phi=0, \tag{1.2a}$$

with a linear part, strongly parabolic of arbitrary order $\nu>\frac{3}{2}$,

$$\operatorname{Re}\hat{P}(i\xi)\geqq \text{Const}\cdot|\xi|^{\nu}, \qquad |\xi|\to\infty. \tag{1.2b}$$

*Received by the editors September 19, 1984, and in revised form May 13, 1985. This research was supported in part by the National Aeronautics and Space Administration under NASA contract NAS1-17070 while the author was in residence at the Institute for Computer Applications in Science and Engineering, NASA Langley Research Center, Hampton, Virginia, 23665. Additional support was provided by National Science Foundation grant DMS85-03294 and Army Research Office grant DAAG29-85-K-0190 while in residence at University of California, Los Angeles.

†School of Mathematical Sciences, Tel-Aviv University, Tel-Aviv 69978, Israel, and Bat-Sheva Foundation Fellow.

[1] G. Sivashinsky, private communication.

Existence and stability results given here, are obtained by modifying Taylor's recipe, [6, p. 96], for treating the existence question in the special case of Burgers equation, $\hat{P}(i\xi)=\xi^2$. According to that recipe, roughly speaking, dissipation is used to compensate nonlinearity, so that short time solutions can be constructed without running into the familiar phenomenon of "loss of derivatives". Coupled with an L^2-decay estimate, short time solutions are then "patched" together, in the large. A study along these lines is carried out in §2 below, where existence and stability questions are treated in connection with the K-S equation. Existence and uniqueness in this case were previously proved by energy methods, see e.g., Aimar and Penel [1], Nicolaenko and Scheurer [3]. The technical details are avoided in §2: these are postponed to §4, all proved by virtue of a single standard estimate on the *linear* dissipative part of the equation, given in §3.

The above study thus suggests itself, with handling arbitrary linear dissipative parts. In §5 we conclude by quoting the corresponding results to such generalized Burgers equations.

2. Existence and stability. We start by putting the K-S equation in a conservative form: we differentiate (1.1), obtaining that the new decayed variable $u\equiv u(x,t;\ \eta)=e^{-\eta t}\partial\phi/\partial x$, $\eta>0$, satisfies

$$\frac{\partial u}{\partial t}+e^{\eta t}\frac{\partial(u^2)}{\partial x}+\eta u+\frac{\partial^2 u}{\partial x^2}+\frac{\partial^4 u}{\partial x^4}=0; \tag{2.1a}$$

a solution for the initial value problem (2.1a) is sought, $u(t)$, $t\geqq 0$, subject to initial condition

$$u(x,t=0)=f(x). \tag{2.1b}$$

Both the pure Cauchy problem, $-\infty<x<\infty$, and the periodic problem, say $-\pi/2\leqq x\leqq\pi/2$, are discussed. We explicitly treat the first infinite case by means of Fourier expansion; the somewhat simpler periodic case can be likewise handled, using Fourier series instead.

If we let $\hat{P}(i\xi)\equiv\hat{P}(i\xi;\ \eta)=\eta-\xi^2+\xi^4$ denote the symbol associated with the spatial linear part of (2.1a) and let $\hat{Q}(\xi,t)\equiv\hat{Q}(\xi,t;\ \eta)=e^{-t\hat{P}(i\xi;\eta)}$ be its transformed solution operator, then by Duhammel's principle (2.1) admits the following integral representation

$$u(t)=Q(t;\ \eta)*f+\int_0^t e^{\eta\tau}\cdot Q(t-\tau;\ \eta)*\frac{\partial}{\partial x}\big(u^2(\tau)\big)\,d\tau. \tag{2.2}$$

Abbreviate the right-hand side of (2.2) by $\mathbf{J}_\eta[u;\ f]$; to simplify notation, we will occasionally suppress the explicit dependence on the initial data, thus writing

$$\mathbf{J}_\eta[u]\equiv\mathbf{J}_\eta[u;\ f]=Q(t;\ \eta)*f+\int_0^t e^{\eta\tau}\cdot Q(t-\tau;\ \eta)*\frac{\partial}{\partial x}\big(u^2(\tau)\big)\,d\tau. \tag{2.3}$$

The question of existence of a solution for (2.1) is now transformed into that of a fixed point solution for $\mathbf{J}_\eta[u]$. Fixing T, $T>0$, we seek a fixed point solution for $\mathbf{J}_\eta[u]$ in $L^\infty([0,T],\ L^2)$, equipped with the standard norm $\|u\|=\sup_{0\leqq t\leqq T}|u(\cdot;\ t)|$.[2] The existence of such a fixed point solution is guaranteed, at least for a short time, as a consequence of

[2] We adopt the notation of single bars to denote spatial norming; for example, $|w|_{H^s}=(\int(1+|\xi|^2)^s|\hat{w}(\xi)|^2\,d\xi)^{1/2}$. Similarly, double bars are reserved to space-time norming; for example, $\|w\|_s=\sup_{0\leqq t\leqq T}|w(\cdot,t)|_{H^s}$. In particular, $|w|=|w|_{H^0}=(\int w^2(x)\,dx)^{1/2}$, $\|w\|=\|w\|_0$.

LEMMA 2.1 (short time contraction). *Given v, w in $L^\infty([0,T], L^2)$ and $\mathbf{J}_\eta[\cdot]=\mathbf{J}_\eta[\cdot;f]$ as in (2.3). Then, there exists a constant $\eta_0 \geqq 0$, such that for $\eta \geqq \eta_0$ we have,*

$$\|\mathbf{J}_\eta[v]-\mathbf{J}_\eta[w]\| \leqq M(T;\eta)\cdot(\|v\|+\|w\|)\cdot\|v-w\|. \tag{2.4a}$$

Here, $M(T;\eta)$ is given by,

$$M(T;\eta)=2e^{\eta T}\cdot T^{1/8}. \tag{2.4b}$$

By virtue of Lemma 2.1 we find

COROLLARY 2.2 (short time boundedness). *Set $T=T_1$, $T_1>0$, such that*

$$4M(T_1;\eta)\cdot|f|<1. \tag{2.5a}$$

Then, for $\eta \geqq \eta_0$ we have,

$$\|\mathbf{J}_\eta^{[n]}[f]\| \leqq 2|f|, \qquad n=0,1,\cdots. \tag{2.5b}$$

Thus, the fixed point iterations, $\mathbf{J}_\eta^{[n]}[f]$ remain inside the origin centered ball of radius $2|f|$. Hence—since by Lemma 2.1 $\mathbf{J}_\eta[\cdot]$ contracts inside that ball, having a Lipschitz constant $4M(T_1;\eta)\cdot|f|<1$—the existence of a fixed point solution for $\mathbf{J}_\eta[u]$ follows, at least for a short time interval, $0 \leqq t \leqq T_1$. Furthermore, *the length of that existence interval, T_1, depends on no higher than the initial L^2-norm.* This latter fact plays a central role in the foregoing analysis; in particular, it enables the local solution just constructed, to be continued to a global one, with the help of

LEMMA 2.3 (large time decay). *Let $u(t;\eta)\equiv u(x,t;\eta)$ be a solution of (2.1). Then, there exists a constant $\eta_0 \geqq 0$, such that for $\eta \geqq \eta_0$ we have*

$$|u(t_2;\eta)| \leqq e^{-(\eta-\eta_0)(t_2-t_1)}\cdot|u(t_1;\eta)|, \qquad 0 \leqq t_1 \leqq t_2 \leqq T. \tag{2.6}$$

Verification of Lemma 2.3 is straightforward: multiplying (2.1a) by $u(x,t;\eta)$, integrating by parts while noting the vanishing contribution of the nonlinear term,[3] we find

$$1/2\,\frac{d}{dt}|u(t)|^2=-\eta|u(t)|^2+\left|\frac{\partial u}{\partial x}(t)\right|^2-\left|\frac{\partial^2 u}{\partial x^2}(t)\right|^2;$$

invoking the Parseval relation, the last equality yields

$$1/2\,\frac{d}{dt}|u(t)|^2 \leqq \max_\xi(-\hat{P}(i\xi;\eta))\cdot|u(t)|^2,$$

and integration finally leads us to (2.6) with $\eta_0=\frac{1}{4}$. We remark that in the periodic case, $-\pi/2 \leqq x \leqq \pi/2$, one can invoke instead Poincaré's inequality,

$$\int_{-\pi/2}^{\pi/2}\left|\frac{\partial u}{\partial x}\right|^2=\int_{-\pi/2}^{\pi/2}\left[\frac{\partial u}{\partial x}-\frac{1}{\pi}\cdot\int_{-\pi/2}^{\pi/2}\frac{\partial u}{\partial x}\right]^2 \leqq \int_{-\pi/2}^{\pi/2}\left|\frac{\partial^2 u}{\partial x^2}\right|^2,$$

leading, in a similar way, to (2.6) with $\eta_0=0$. Observe that in general the exponential growth bound, η_0, may depend on the period.

To conclude the existence of solution in the large, we now fix η, $\eta \geqq \eta_0$, with appropriately chosen η_0 in either the finite or infinite case; then, short time

[3] With the infinite pure Cauchy problem, $u(x,t)$ is required to vanish at $x=\pm\infty$, indeed, $|u(t)|_{H^1}<\infty$ according to Theorem 2.6 below.

solutions—constructed according to Lemma 2.1—can be successively "patched" together, over time intervals which—according to Lemma 2.3—are of a *fixed* (non-shrinking) length T_1. Integrating, we obtain a global solution for the K-S equation, $\phi=\phi(x,t)$; the solution so obtained is—up to integration factor—unique. Thus we finally arrive at

THEOREM 2.4 (existence). *The* K-S *equation* (1.1), *with prescribed initial data* $\phi(t=0)$ *in* H^1, *admits a unique solution*, $\phi=\phi(x,t)$, *which satisfies*,

$$\left|\frac{\partial\phi}{\partial x}(t)\right|\leqq e^{\eta_0 T}\cdot\left|\frac{\partial\phi}{\partial x}(t=0)\right|,\qquad 0\leqq t\leqq T<\infty. \tag{2.7}$$

In fact, $\phi(t)$, $t\geqq 0$, belongs to H^1: a further L^2 estimate needed here, is discussed in §4 below.

The global solution referred to in Theorem 2.4, is constructed by patching together short time solutions, using a single L^2 a priori estimate. Such a patching procedure differs from existence proofs via standard energy methods, e.g., [1], [3], where higher a priori estimates are called for. Instead, we rely here on having a derivative-free Lipschitz contraction factor, so that short time solutions can be constructed, without running into the familiar phenomenon of "loss of derivatives". We note that solving the integrodifferential equation (2.2) by fixed point iterations results in the existence of a solution satisfying the original *differential* equation (2.1), in a *weak* sense. Concerning the existence of such a solution under a stronger topology, one observes that (2.1a) contains two destabilizing sources: the focusing effect ("loss of derivatives") caused by the nonlinear term, and the exponential divergence of the second order dissipative term. It is the balance of these two terms by the fourth-order dissipation, which leads us to the important derivative-free Lipschitz contraction factor in this case. Making a finer study of that balance, we are able to conclude that the solution constructed above is, in fact, smooth enough to be interpreted as a classical one. To this end, we sharpen Lemma 2.1, stating

LEMMA 2.5 (short time contraction). *Given* v, w *in* $L^\infty([0,T],\ H^s)$ $s\geqq 0$, *and* $\mathbf{J}_\eta[\cdot]=\mathbf{J}_\eta[\cdot;\ f]$ *as in* (2.3). *Then*, *there exists a constant* $\eta_0\geqq 0$, *such that for* $\eta\geqq\eta_0$ *we have*,

$$\|\mathbf{J}_\eta[v]-\mathbf{J}_\eta[w]\|_{s+2}\leqq 2^s\cdot M(T;\ \eta)\cdot(\|v\|_s+\|w\|_s)\cdot\|v-w\|_s. \tag{2.8}$$

Thus, each fixed point iteration gives us a smoother *correction*. In particular, setting s to be zero, we find on account of Corollary 2.2 that $\{\mathbf{J}_\eta^{[n]}[f]\}_{n\geqq 0}$ form a Cauchy sequence in the $L^\infty([0,T_1],\ H^2)$—origin centered ball of radius $2|f|$. Hence, the fixed point iterations $\mathbf{J}_\eta^{[n]}[f]$ converge to a unique, short time solution, $u=u(x,t)$ in $L^\infty([0,T_1],\ H^2)$. Thanks to the L^2-decay estimate in Lemma 2.3, such short time solutions can be patched in the large as before, integrated once and yielding

THEOREM 2.6 (existence). *The* K-S *equation* (1.1), *with prescribed initial data* $\phi(t=0)$ *in* H^3, *admits a unique solution*, $\phi=\phi(x,t)$, *which satisfies*,

$$\left|\frac{\partial\phi}{\partial x}(t)\right|_{H^2}\leqq\frac{5}{4}e^{\alpha T}\cdot\left|\frac{\partial\phi}{\partial x}(t=0)\right|_{H^2},\qquad 0\leqq t\leqq T\leqq\infty. \tag{2.9}$$

Finally, we turn to examine the question of stability: allowing the initial data to vary as well, we have the final extension to the short time contraction lemma, which now reads

LEMMA 2.7 (short time contraction). *Given v, w in $L^\infty([0,T],\ H^s)$ with $f=v(t=0)$, $g=w(t=0)$ in H^{s+2}. Then, there exists a constant $\eta_0\geqq 0$ such that for $\eta\geqq\eta_0$ we have,*

$$\|\mathbf{J}_\eta[v;\ f]-\mathbf{J}_\eta[w;\ g]\|_{s+2} \leqq |f-g|_{H^{s+2}}+2^s\cdot M(t;\ \eta)\cdot(\|v\|_s+\|w\|_s)\cdot\|v-w\|_s. \tag{2.10}$$

Now let $v(t)=\mathbf{J}_\eta[v(t);\ v(t=0)]$, $w(t)=\mathbf{J}_\eta[w(t);\ w(t=0)]$ be two different fixed point solutions of (2.1a), whose initial data $f=v(t=0)$ and $g=w(t=0)$ are assumed to be in H^2; according to Theorem 2.6, $u(t)$ and $v(t)$ belong to H^2 later on, $t\geqq 0$, and as a consequence of Lemma 2.7 with $s=0$, we have short time stability

$$|v(t)-w(t)|_{H^2}\leqq\frac{1}{1-M(T_1;\ \eta)\cdot(|f|+|g|)}|v(t=0)-w(t=0)|_{H^2},\qquad 0\leqq t\leqq T_1.$$

Successive application of the last inequality yields the desired stability result, which we state as our final

THEOREM 2.8 (stability). *Let ϕ, ψ be two different solutions of the* K-S *equation* (1.1), *with initial data $\phi(t=0)$, $\psi(t=0)$ lying in H^3. Then, there exist constants C and $\beta\geqq 0$ (both may depend on $|(\partial\phi/\partial x)(t=0)|+|(\partial\psi/\partial x)(t=0)|$), such that the following estimate holds*:

$$\left|\frac{\partial\phi}{\partial x}(t)-\frac{\partial\psi}{\partial x}(t)\right|_{H^2}\leqq C\cdot e^{\beta t}\cdot\left|\frac{\partial\phi}{\partial x}(t=0)-\frac{\partial\psi}{\partial x}(t=0)\right|_{H^2},\qquad 0\leqq t\leqq T\leqq\infty. \tag{2.11}$$

3. An estimate on the dissipative kernel. The following classical estimate is in the heart of the matter.

LEMMA 3.1. *Given $\omega(x)$ in W_p^m, $1\leqq p\leqq 2$, and real r, $r\geqq\frac{1}{2}-1/p$. Then, there exist constants, $C=C_{p,r}$ and $\eta_0\geqq 0$, such that for $\eta\geqq\eta_0$ we have,*

$$|Q(t;\ \eta)*\omega|_{H^{m+r}}\leqq C\cdot e^{-(\eta-\eta_0)t}\cdot t^{-(r-1/2+1/p)/4}\cdot|\omega|_{W_p^m}. \tag{3.1}$$

Remark. We adopt here the standard notation, W_p^m, to denote the L^p-type Sobolev space of order m, consisting of those functions whose derivatives up to order m belong to L^p. (Although not specifically referred to, a fractional Sobolev space with nonintegral m should be interpreted as a Besov space: to comply with notation, we therefore restrict attention to integral orders, with the understanding that final results can be interpolated into Besov space.)

For completeness, we include here a short calculation verifying (3.1): setting $\mu=p/(2-p)$ and letting μ' be its conjugate, $1/\mu+1/\mu'=1$; then the Hölder inequality yields

$$|Q(t;\ \eta)*\omega|_{H^{m+r}}\leqq\left[\int_{-\infty}^{\infty}\left(1+|\xi|^2\right)^{\mu r}e^{-2\mu t(\eta-\xi^2+\xi^4)}d\xi\right]^{1/2\mu} \times\left[\int_{-\infty}^{\infty}\left(1+|\xi|^2\right)^{\mu' m}|\hat{\omega}(\xi)|^{2\mu'}d\xi\right]^{1/2\mu'}. \tag{3.2}$$

Since by the Hausdorff–Young inequality the Fourier transform is of type $(2\mu',\ (2\mu')'=p)$, the second factor on the right of (3.2), $|\hat{\omega}|_{W_{2\mu'}^m}$, does not exceed

$$\left[\int_{-\infty}^{\infty}\left(1+|\xi|^2\right)^{\mu' m}|\hat{\omega}(\xi)|^{2\mu'}d\xi\right]^{1/2\mu'}\leqq(2\pi)^{1/2-1/p}\cdot|\omega|_{W_p^m}. \tag{3.3}$$

Next, we split the first factor on the right of (3.2),

$$e^{-\eta t}\left[\int_{-\infty}^{\infty}\left(1+|\xi|^2\right)^{\mu r}e^{-2\mu t(\xi^4-\xi^2)}d\xi\right]^{1/2\mu}=e^{-\eta t}\left[\int_{|\xi|\leq\sqrt{2}}\cdots+\int_{|\xi|>\sqrt{2}}\cdots\right]^{1/2\mu};$$

the first of the two integrals admits a pessimistic bound of

$$\int_{|\xi|\leq\sqrt{2}}\left(1+|\xi|^2\right)^{\mu r}e^{-2\mu t(\xi^4-\xi^2)}d\xi\leq 2\sqrt{2}\,3^{\mu r}e^{\mu t/2},$$

while the second one is estimated by

$$\int_{|\xi|>\sqrt{2}}\left(1+|\xi|^2\right)^{\mu r}e^{-2\mu t(\xi^4-\xi^2)}d\xi$$

$$\leq 2^{\mu r+1}\int_{\xi=0}^{\infty}\xi^{2\mu r}e^{-\mu t\xi^4}d\xi=2^{\mu r-1}\Gamma\left(\frac{2\mu r+1}{4}\right)(\mu t)^{-(2\mu r+1)/4}.$$

Added together, we find that the first factor on the right of (3.2), does not exceed

$$\left[\int_{-\infty}^{\infty}\left(1+|\xi|^2\right)^{\mu r}e^{-2\mu t(\eta-\xi^2+\xi^4)}d\xi\right]^{1/2\mu} \tag{3.4}$$

$$\leq B_{p,r}\cdot e^{-(\eta-\eta_0)t}\cdot t^{-(2\mu r+1)/8\mu},\qquad \eta_0=\frac{1}{4},$$

with Stirling's formula giving us a bound of

$$B_{p,r}=(4\pi e)^{1/2\mu}3^{r/2}\left(r+\frac{1}{p}\right)^{r-1/2+1/p}.$$

Recalling that $(2\mu')'=p$, (3.2), (3.3) and (3.4) yield the required estimate (3.1) with $C_{p,r}=(2\pi)^{1/2-1/p}B_{p,r}$.

Remark 1. In the *infinite* case under consideration, an exponential growth bound, $\eta_0=\frac{1}{4}$, was found. In general, η_0 may depend on the period, in the spirit of an earlier remark; for example, $\eta_0=0$, in the π-periodic case.

Remark 2. For future reference, we quote here the constants $C_{p,r}$ in two special cases: as can be readily verified, $C_{2,0}=1$ (indeed, such an estimate also follows by a straightforward integration by parts, essentially contained in the verification of Lemma 2.3 above); also, by sharpening the above pessimistic bounds, one finds $C_{1,3}<8$.

4. Proof of main results. We first study the operator $\mathbf{J}_\eta[\cdot;\ \cdot]$ introduced in (2.3), whose fixed point solutions are sought. Equipped with Lemma 3.1, we are able to derive the following summary stability estimate

$$\text{(S)}\quad \left|\mathbf{J}_\eta[v(t);\ f]-\mathbf{J}_\eta[w(t);\ g]\right|_{H^{s+2}}\leq e^{-(\eta-\eta_0)t}\cdot|f-g|_{H^{s+2}}$$

$$+2^{s+1}\cdot e^{\eta t}\cdot t^{1/8}\cdot\sup_{0\leq\tau\leq t}|v(\tau)+w(\tau)|_{H^s}\cdot\sup_{0\leq\tau\leq t}|v(\tau)-w(\tau)|_{H^s}.$$

To verify (S)—assuming the quantities on the right are finite and $\eta\geq\eta_0$—we consider the difference

$$\mathbf{J}_\eta[v(t);\ f]-\mathbf{J}_\eta[w(t);\ g]=Q(t;\ \eta)*(f-g)$$

$$+\int_0^t e^{\eta\tau}\cdot Q(t-\tau;\ \eta)*\frac{\partial}{\partial x}\left(v^2(\tau)-w^2(\tau)\right)d\tau,$$

so that after taking norms on both sides we have

$$\left|\mathbf{J}_\eta[v(t);\, f]-\mathbf{J}_\eta[w(t);\, g]\right|_{H^{s+2}} \leqq |Q(t;\,\eta)*(f-g)|_{H^{s+2}} + \int_0^t e^{\eta\tau}\cdot\left|Q(t-\tau;\,\eta)*\frac{\partial}{\partial x}\left(v^2(\tau)-w^2(\tau)\right)\right|_{H^{s+2}} d\tau.$$

Now applying Lemma 3.1 with respect to both terms on the right of the last inequality: the first term with $(r,p,m)=(0,2,s+2)$, and the second one with $(r,p,m)=(3,1,s-1)$; recalling the earlier quoted constants $C_{2,0}=1$ and $C_{1,3}<8$, we find

$$\left|\mathbf{J}_\eta[v(t);\, f]-\mathbf{J}_\eta[w(t);\, g]\right|_{H^{s+2}} \leqq e^{-(\eta-\eta_0)t}\cdot|f-g|_{H^{s+2}} + 8\cdot\int_0^t e^{\eta\tau}\cdot e^{-(\eta-\eta_0)(t-\tau)}\cdot(t-\tau)^{-7/8}\cdot\left|\frac{\partial}{\partial x}\left(v^2(\tau)-w^2(\tau)\right)\right|_{W_1^{s-1}} d\tau.$$

The last integral bounds the interaction between the linear dissipative part of the equation, and the nonlinear differentiated quadratic term; the loss of derivative due to the latter is compensated here by dissipation, weighted with the L^1 topology. In order to return to the usual L^2 setup, we apply the Leibniz rule and Cauchy–Schwarz inequality to find

$$\left|\frac{\partial}{\partial x}\left(v^2(\tau)-w^2(\tau)\right)\right|_{W_1^{s-1}} \leqq 2^{s+1}\cdot|v(\tau)+w(\tau)|_{H^s}\cdot|v(\tau)-w(\tau)|_{H^s}.$$

Inserted into the last integral and carrying out the integration, we end up with the required estimate (S).

We now turn to prove the results in §2, starting with:

Short time contractions (Lemma 2.1, Lemma 2.5, Lemma 2.7). Taking supremum over both sides of the (S) estimate with varying t, $0\leqq t\leqq T$, and equipped with the notation of

$$M(T;\,\eta)=2e^{\eta T}\cdot T^{1/8}$$

in (2.4b), we find

$$\left\|\mathbf{J}_\eta[v;\, f]-\mathbf{J}_\eta[w;\, g]\right\|_{s+2} \leqq |f-g|_{H^{s+2}}+2^s\cdot M(T;\,\eta)\cdot\left(\|v\|_s+\|w\|_s\right)\cdot\|v-w\|_s,$$

so that Lemma 2.7 follows. Taking the special case $f=g$ proves Lemma 2.5, and further setting $s=0$, yields Lemma 2.1,

$$\left\|\mathbf{J}_\eta[v]-\mathbf{J}_\eta[w]\right\| \leqq \left\|\mathbf{J}_\eta[v]-\mathbf{J}_\eta[w]\right\|_2 \leqq M(T;\,\eta)\cdot\left(\|v\|+\|w\|\right)\cdot\|v-w\|.$$

(Observe that in the case of Lemma 2.1, where no gain of derivatives is involved, one can in fact improve the contraction factor $M(T;\,\eta)$ to be $\frac{2}{7}e^{\eta T}T^{7/8}$.)

An immediate consequence of Lemma 2.1 is the following:

Short time boundedness (Corollary 2.2). Setting $v=\mathbf{J}_\eta^{[n-1]}[f]$ and $w=0$ in Lemma 2.1, we find

$$\left\|\mathbf{J}_\eta\left[\mathbf{J}_\eta^{[n-1]}(f)\right]\right\| \leqq \left\|\mathbf{J}_\eta[v;\, f]-\mathbf{J}_\eta[w=0;\, f]\right\|+\left\|\mathbf{J}_\eta[w=0;\, f]\right\| \leqq M(T;\,\eta)\cdot\left\|\mathbf{J}_\eta^{[n-1]}[f]\right\|^2+\|Q(t;\,\eta)*f\|.$$

We now consider a temporal interval of length T_1 such that $4M(T_1;\ \eta)\cdot|f|<1$: assuming $\|\mathbf{J}_\eta^{[n-1]}[f]\|\leqq 2|f|$ in that interval, then together with Lemma 3.1 taking $(r,p,m)=(0,2,0)$, we obtain

$$\left\|\mathbf{J}_\eta^{[n]}[f]\right\|\leqq 4M(T_1;\ \eta)|f|\cdot|f|+|f|\leqq 2|f|,$$

and Corollary 2.2 follows by induction.

Owing to the last two results in the small, one may construct fixed point solutions, $u(t)$, as local solutions over time intervals $[T_N,T_{N+1}]$ $N=0,1,2,\cdots$, such that $4M(T_{N+1}-T_N;\ \eta)\cdot|u(T_N)|<1$. Thanks to the large L^2-estimate in Lemma 2.3, the local solutions just constructed can be patched in the large, over *fixed* length time intervals, $T_N=NT_1$, $N=0,1,\cdots$, obtaining

Existence. (Theorem 2.4, Theorem 2.6). Given the initial data $\phi(t=0)$ in H^1, we set $f=(\partial\phi/\partial x)(t=0)$ for the initial value problem (2.1); let $u(t)$, $t\geqq 0$, be its global solution, constructed according to the above recipe. Integrated once, we obtain a solution for the K-S equation, $\phi(x,t)=\int^x u(\xi,t)\,d\xi$, which satisfies—choosing $\eta=\eta_0$ in Lemma 2.3—

$$\left|\frac{\partial\phi}{\partial x}(t)\right|\leqq e^{\eta_0 T}\cdot\left|\frac{\partial\phi}{\partial x}(t=0)\right|,\qquad 0\leqq t\leqq T.$$

This proves Theorem 2.4. In order to show that $u=\partial\phi/\partial x$ possesses a certain degree of smoothness, at least that of the initial data, we appeal to the short time contraction estimate in Lemma 2.5 with $s=0$:

$$\left\|\mathbf{J}_\eta[v]-\mathbf{J}_\eta[w]\right\|_2\leqq M(T;\ \eta)\cdot(\|v\|+\|w\|)\cdot\|v-w\|.$$

Consider first the time interval $[0,T=T_1]$ and let $u=\mathbf{J}_\eta[u]$ the fixed point solution there; choosing $v=u$ and $w=0$, we find

$$\|u\|_2=\left\|\mathbf{J}_\eta[u]\right\|_2\leqq M(T_1;\ \eta)\|u\|^2+\|Q*f\|_2.$$

Using Lemma 2.3 and Lemma 3.1 with $(r,p,m)=(0,2,2)$, we end up with

$$\|u\|_2\leqq M(T_1;\ \eta)|f|^2+|f|_2^2\leqq\frac{5}{4}|f|_2.$$

Successive application of the last inequality over the accumulated patching intervals, implies

$$|u(t;\ \eta)|_{H^2}\leqq\left(\frac{5}{4}\right)^{t/T_1+1}\cdot|f|_{H^2}.$$

Choosing $\eta=\eta_0$, Theorem 2.6 now follows with $\alpha=\eta_0+\ln(\frac{5}{4})$,

$$\left|\frac{\partial\phi}{\partial x}(t)\right|_{H^2}\leqq\frac{5}{4}e^{\alpha T}\cdot\left|\frac{\partial\phi}{\partial x}(t=0)\right|_{H^2},\qquad 0\leqq t\leqq T<\infty.$$

Remark. We note that the above solution $\phi=\phi(x,t)$ lies, in fact, in the same Sobolev space the initial data belong to, H^s, $0\leqq s\leqq 2$. This follows from a complementing L^2-estimate which we now derive: multiplying (1.1) by ϕ and integrating by parts, we find

$$\frac{1}{2}\frac{d}{dt}|\phi(t)|^2\leqq\left|\frac{\partial\phi}{\partial x}(t)\right|^2-\left|\frac{\partial^2\phi}{\partial x^2}(t)\right|^2+|\phi(t)|_{L^\infty}\cdot\left|\frac{\partial\phi}{\partial x}(t)\right|^2.$$

We interpolate in a somewhat nonstandard way, $|\phi|_{L^\infty} \leq \varepsilon|\phi| + C \cdot \varepsilon^{-1} \cdot |\phi_x|$, so that by appropriately choosing $\varepsilon = \gamma \cdot |(\partial\phi/\partial x)(t)|^{-2}$, the last inequality implies

$$\frac{1}{2}\frac{d}{dt}|\phi(t)|^2 \leq \gamma \cdot |\phi(t)|^2 + \gamma^{-1}K,$$

with $K = K(|(\partial\phi/\partial x)(t)|)$. Thanks to Lemma 2.3, we can control

$$K\left(\left|\frac{\partial\phi}{\partial x}(t)\right|\right) \leq K\left(\left|\frac{\partial\phi}{\partial x}(t=0)\right|\right),$$

and L^2-boundedness now follows

$$|\phi(t)| \leq e^{\gamma t} \cdot \left[|\phi(t=0)| + \gamma^{-1} \cdot K\left(\left|\frac{\partial\phi}{\partial x}(t=0)\right|\right)\right],$$

with arbitrarily small exponential growth factor γ, $\gamma > 0$. Regarding the periodic case, $-\pi/2 \leq x \leq \pi/2$, one may subtract the average

$$\bar{\phi}(t) = \frac{1}{\pi}\int_{-\pi/2}^{\pi/2} \phi(x,t)\,dx,$$

so that by invoking Poincaré's inequality for $\phi(t) - \bar{\phi}(t)$ rather than interpolating, we find

$$|\phi(t) - \bar{\phi}(t)| \leq |\phi(t=0) - \bar{\phi}(t=0)| + K\left(\left|\frac{\partial\phi}{\partial x}(t=0)\right|\right) \cdot t^{1/2}.$$

5. A generalized Burgers equation. The results of the last sections were so organized, in order to emphasize that the only a priori estimate required for the proofs, concerns the linear dissipative part of the equation, see Lemma 3.1. Hence, the following generalization can be easily worked out.

We consider the generalized Burgers equation

$$\frac{\partial u}{\partial t} + \frac{\partial(u^2)}{\partial x} + P\left(\frac{\partial}{\partial x}\right)u = 0 \tag{5.1a}$$

whose linear part, $\partial/\partial t + P(\partial/\partial x)$, is assumed strongly parabolic of order ν,

$$\operatorname{Re}\hat{P}(i\xi) \geq \text{Const} \cdot |\xi|^\nu, \qquad |\xi| \to \infty. \tag{5.1b}$$

Regarding the corresponding kernel, $\hat{Q}(t;\ \eta) = e^{-t(\eta + \hat{P}(i\xi))}$, we have, in analogy with Lemma 3.1,

$$|Q(t;\ \eta) * \omega|_{H^{m+r}} \leq C \cdot e^{-(\eta - \eta_0)t} \cdot t^{-(r - 1/2 + 1/p)/\nu} \cdot |\omega|_{W_p^m}. \tag{5.2}$$

In particular, considering $Q(t;\ \eta)$ operating from L^1 to H^{1+s}, it is found to have an operator norm with an *integrable* singularity, $t^{-(s+3/2)/\nu}$, provided $s < \nu - \frac{3}{2}$. Arguments similar to those introduced in §2, then lead us to

THEOREM 5.1. *Let u, v be two different solutions of the generalized Burgers equation* (5.1), *with initial data lying in H^s, $s < \nu - 3/2$. Then, there exist constants, C and $\beta \geq 0$* (*both may depend on $|u(t=0)| + |v(t=0)|$*), *such that the following estimate holds*:

$$|u(t) - v(t)|_{H^s} \leq C \cdot e^{\beta t} \cdot |u(t=0) - v(t=0)|_{H^s}. \tag{5.3}$$

We end up noting that the above recipe suggests itself, in studying the all important question regarding the long-time behavior of solutions for (5.1).

Remark. The special case $P(\partial/\partial x)=(-\partial^2/\partial x^2)^{\nu/2}$ can be considered as a one-dimensional degenerate case of the formal d-dimensional Navier–Stokes equations; global regularity in the latter case follows with dissipativity of order $\nu>1+d/2$, (see, e.g., Rose and Sulem, J. de Physique, 39 (1978), pp. 441–484). In either way, one finds $\nu=\frac{3}{2}$ as the critical order of dissipativity which guarantees regularity in the one-dimensional case, $d=1$.

REFERENCES

[1] M. T. AIMAR AND P. PENEL, *Résultats d'existence et d'unicité du modèle de diffusion nonlineaire de G. I. Sivashinsky*, Université de TOULON et du VAR,Preprint, 1982.

[2] Y. KURAMOTO, *Instability and turbulence of wave fronts in reaction-diffusion systems*, Progr. Theoret. Phys., 63 (1980), pp. 1885–1903.

[3] B. NICOLAENKO AND B. SCHEURER, *Remarks on the Kuramoto–Sivashinsky equation*, Proc. Conference on Fronts, Interfaces and Patterns, Physica D, 12D (1984), pp. 391–395.

[4] G. SIVASHINSKY, *Nonlinear analysis of hydrodynamic instability in laminar flames, Part* I, *Derivation of basic equations*, Acta Astronaut., 4 (1977), pp. 1117–1206.

[5] T. SHLANG AND G. SIVAHSINSKY, *Irregular flow of a liquid film down a vertical column*, J. de Physique, 43 (1982), pp. 459–466.

[6] M. TAYLOR, *Pseudodifferential Operators*, Princeton Univ. Press, Princeton, NJ, 1981.

SIAM J. MATH. ANAL.
Vol. 17, No. 4, July 1986

010

HOPF BIFURCATION IN TWO-COMPONENT FLOW*

M. RENARDY† AND D. D. JOSEPH‡

Abstract. The stability of viscosity-stratifed bicomponent flow has been studied by long wave asymptotics, by short wave asymptotics and numerically. These studies have shown that interfacial instabilities arise from the viscosity difference between the two fluids. If the surface tension between the fluids is nonzero, then Hopf type bifurcatinos leading to traveling interfacial waves are expected. In this paper, we prove a rigorous theorem establishing the existence of bifurcating solutions of this nature.

Key words. two-component flow, Hopf bifurcation

AMS(MOS) subject classifications. Primary 35B32, 35Q10, 76E05, 76V05

1. Introduction. The stability of two-component parallel shear flows has been analyzed by long-wave asymptotics [5], [16], short-wave asymptotics [6], [13] and numerically [11], [13]. These studies show that, if the fluids have different viscosities, then instabilities can arise at all Reynolds numbers.

This raises the question of possible alternative flow patterns which might be stable. Yih [16] has conjectured that wavy interfaces might develop. The analysis of Hooper and Boyd [6] reveals a crucial difference between the cases of zero and nonzero surface tension between the fluids. If the surface tension is zero, then sufficiently short waves are always unstable, i.e., there is an infinite number of unstable modes. This situation is very much unlike the usual problems of bifurcation theory, and we believe it is possible that no smooth interface, steady or unsteady, would be stable in this situation. (In reality, of course, the surface tension is not zero, but there will be instability for very short waves when the surface tension is small and the Reynolds number is large. We think that this instability mechanism may be relevant in the formation of emulsions.)

In the case of nonzero surface tension, however, one can establish a bifurcation theorem. If the bifurcation turns out supercritical, this provides a basis for Yih's conjecture. Whether or not the bifurcation is supercritical will in general have to be decided by a numerical calculation. To our knowledge, such calculations have not yet been done. The computations referred to above concern only the eigenvalues of the linearized problems. For the sake of simplicity, we confine attention to plane Couette flow, but it is clear that similar techniques can be applied to more complicated geometries such as concentric flow in pipes or between rotating cylinders. We consider plane Couette flow of two fluids with equal density, but different viscosities, and an interface parallel to the plates. Periodic boundary conditions are imposed in the streamwise direction. Evidently, this configuration is stable at rest. If there is a flow,

* Received by the editors October 8, 1984, and in revised form April 12, 1985.

† Department of Mathematics and Mathematics Research Center, University of Wisconsin-Madison, Madison, Wisconsin 53706. The work of this author was supported by the U. S. Army under contract DAAG29-80-C-0041. This material is based upon work supported by the National Science Foundation under grants MCS-8210950 and MCS-8215064.

‡ Department of Aerospace Engineering, University of Minnesota, Minneapolis, Minnesota 55455. The work of this author was supported by the U. S. Army under contract DAAG29-82-K-0029 and the Fluid Mechanics branch of the National Science Foundation.

however, then instabilities can develop [6], [16]. Since, however, surface tension will damp short waves, there can be only a finite number of unstable modes. Generically, as the flow rate is increased, one specific mode will be the first to become unstable. Since the eigenvalues are complex, one expects a bifurcation of the Hopf type [7], leading to traveling interfacial waves.

The Hopf bifurcation thoerem in infinite dimensions [4], [8]–[10], [14] relies on coercive estimates for the linearized equations. For one-component free surface flows such estimates were derived by Beale [2], [3], and we shall, in §3, derive analogous estimates for two-component flow. Our proof differs from Beale's and is slightly simpler. Using these coercive estimates, we can then establish a bifurcation theorem in §4. In §§5–8, we outline an algorithm for the computation of bifurcating solutions.

2. Formulation of the problem. We consider two-dimensional flow of two fluids with different viscosities and equal densities between parallel plates; see Fig. 1. The motion in each fluid is described by the Navier–Stokes equations:

$$\left.\begin{aligned} &\rho(\dot{\mathbf{u}}+(\mathbf{u}\cdot\nabla)\mathbf{u})=\eta_1\Delta\mathbf{u}-\nabla p,\\ &\nabla\cdot\mathbf{u}=0,\end{aligned}\right\} \qquad 0<y<h(x), \tag{2.1}$$

$$\left.\begin{aligned} &\rho(\dot{\mathbf{v}}+(\mathbf{v}\cdot\nabla)\mathbf{v})=\eta_2\Delta\mathbf{v}-\nabla q,\\ &\nabla\cdot\mathbf{v}=0,\end{aligned}\right\} \qquad h(x)<y<1. \tag{2.2}$$

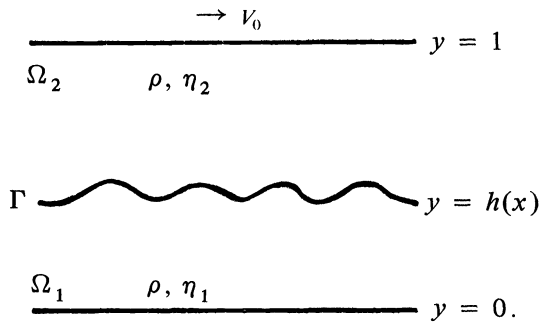


FIG. 1

We have no slip conditions at the walls:

$$\mathbf{u}=0 \qquad \text{at } y=0, \tag{2.3}$$

$$\mathbf{v}=(V_0,0) \qquad \text{at } y=1. \tag{2.4}$$

Across the interface, there must be continuity of velocity,

$$\mathbf{u}=\mathbf{v} \quad \text{at } y=h(x), \tag{2.5}$$

continuity of shear stress

$$(2.6)\quad \eta_1\left\{(1-h'^2)\left(\frac{\partial u_2}{\partial x}+\frac{\partial u_1}{\partial y}\right)+2h'\left(\frac{\partial u_2}{\partial y}-\frac{\partial u_1}{\partial x}\right)\right\}$$
$$=\eta_2\left\{(1-h'^2)\left(\frac{\partial v_2}{\partial x}+\frac{\partial v_1}{\partial y}\right)+2h'\left(\frac{\partial v_2}{\partial y}-\frac{\partial v_1}{\partial x}\right)\right\} \quad \text{at } y=h(x),$$

and balance of the normal stress difference by surface tension

$$(2.7)\quad 2\eta_1\frac{\partial u_2}{\partial y}-p-2h'\eta_1\left(\frac{\partial u_2}{\partial x}+\frac{\partial u_1}{\partial y}\right)+h'^2\left(2\eta_1\frac{\partial u_1}{\partial x}-p\right)$$
$$=2\eta_2\frac{\partial v_2}{\partial y}-q-2h'\eta_2\left(\frac{\partial v_2}{\partial x}+\frac{\partial v_1}{\partial y}\right)+h'^2\left(2\eta_2\frac{\partial v_1}{\partial x}-q\right)$$
$$+T\frac{h''}{(1+h'^2)^{1/2}} \quad \text{at } y=h(x).$$

Here T is the surface tension parameter. Finally, we have the kinematic boundary condition

$$(2.8)\qquad \dot{h}+u_1h'=u_2.$$

We are interested in solutions to (2.1)–(2.8) which have a given period L in the x-direction and are periodic in t. We denote by Ω_1 the set $\{(x,y)\,|\,0\leqq x|L,\ 0\leqq y\leqq h(x)\}$, by Ω_2 the set $\{(x,y)\,|\,0\leqq x\leqq L, h(x)\leqq y\leqq 1\}$ and by Γ the interface $\{(x,y)\,|\,0\leqq x\leqq L, y=h(x)\}$. The spaces $H^k(\Omega_1)$, $H^k(\Omega_2)$, $H^k(\Gamma)$ consist of those functions which have k square integrable derivatives and satisfy periodic boundary conditions in the x-direction.

3. The linearized problem. In this chapter, we obtain coercive estimates on the linear problem, which we shall need later. We put $V_0=0$ and linearize (2.1)–(2.8) at the rest state $\mathbf{u}=\mathbf{v}=0$, with a flat interface $h=h_0$. We include inhomogeneous terms in (2.1), (2.2), (2.6) and (2.7). The goal of this section is to derive estimates for a resolvent operator, i.e. the time dependence is assumed to be exponential, and $\partial/\partial t$ can be replaced by a constant factor λ. This leads to the problem:

$$(3.1)\quad \left.\begin{aligned}\lambda\rho\mathbf{u}&=\eta_1\Delta\mathbf{u}-\nabla p+\mathbf{f}_1,\\ \nabla\cdot\mathbf{u}&=0,\end{aligned}\right\} \qquad 0<y<h_0$$

$$(3.2)\quad \left.\begin{aligned}\lambda\rho\mathbf{v}&=\eta_2\Delta\mathbf{v}-\nabla q+\mathbf{f}_2\\ \nabla\cdot\mathbf{v}&=0,\end{aligned}\right\} \qquad h_0<y<1,$$

$$(3.3)\quad \mathbf{u}=0, \qquad y=0,$$

$$(3.4)\quad \mathbf{v}=0, \qquad y=1,$$

$$(3.5)\quad \mathbf{u}=\mathbf{v}, \qquad y=h_0,$$

$$(3.6)\quad \eta_1\left(\frac{\partial u_2}{\partial x}+\frac{\partial u_1}{\partial y}\right)-\eta_2\left(\frac{\partial v_2}{\partial x}+\frac{\partial v_1}{\partial y}\right)=f_3, \qquad y=h_0,$$

$$(3.7)\quad 2\eta_1\frac{\partial u_2}{\partial y}-p-2\eta_2\frac{\partial v_2}{\partial y}+q-Th''=f_4, \qquad y=h_0,$$

$$(3.8)\quad \lambda h=u_2, \quad y=h_0.$$

We seek solutions periodic in x with period L, and of course the same periodicity is assumed for $\mathbf{f}_1, \mathbf{f}_2$, f_3 and f_4. Our goal is the following estimate:

THEOREM 3.1. *Let* $\gamma > 0$. *For* $\mathrm{Re}\lambda \geqq \gamma$, *the following estimate holds for solutions of* (3.1)–(3.8):

$$
\begin{aligned}
(3.9)\qquad & \|\mathbf{u}\|_{H^2(\Omega_1)} + \|\mathbf{v}\|_{H^2(\Omega_2)} + \|p\|_{H^1(\Omega_1)} + \|q\|_{H^1(\Omega_2)} + |\lambda|\|\mathbf{u}\|_{L^2(\Omega_1)} \\
& \quad + |\lambda|\|\mathbf{v}\|_{L^2(\Omega_2)} + \|h\|_{H^{5/2}(\Gamma)} + |\lambda|\|h\|_{H^{3/2}(\Gamma)} \\
& \quad \leqq C\left[\|\mathbf{f}_1\|_{L^2(\Omega_1)} + \|\mathbf{f}_2\|_{L^2(\Omega_2)} + \|f_3\|_{H^{1/2}(\Gamma)} + |\lambda|^{1/4}\|f_3\|_{L^2(\Gamma)} + \|f_4\|_{H^{1/2}(\Gamma)}\right].
\end{aligned}
$$

(*Here C can depend on* γ *but not on* λ.)

Proof. We multiply (3.1) by $\bar{\mathbf{u}}$ (the complex conjugate of $\mathbf{u}$), and (3.2) by $\bar{\mathbf{v}}$, add them and integrate over the domain. Integrating by parts, and using the boundary and interface conditions, we find

$$
\begin{aligned}
(3.10)\qquad & \lambda\rho\left(\|\mathbf{u}\|_{L^2}^2 + \|\mathbf{v}\|_{L^2}^2\right) + \frac{1}{2}\eta_1\langle\mathbf{u},\mathbf{u}\rangle + \frac{1}{2}\eta_2\langle\mathbf{v},\mathbf{v}\rangle + \bar{\lambda}T\|h'\|_{L^2}^2 \\
& \quad = (\mathbf{f}_1, \bar{\mathbf{u}})_{L^2} + (\mathbf{f}_2, \bar{\mathbf{v}})_{L^2} + \int_0^L f_3\bar{u}_1(y = h_0)\,dx + \int_0^L f_4\bar{u}_2(y = h_0)\,dx.
\end{aligned}
$$

Here $\langle\mathbf{u},\mathbf{u}\rangle = \int (\nabla\mathbf{u} + (\nabla\mathbf{u})^T) : (\nabla\bar{\mathbf{u}} + (\nabla\bar{\mathbf{u}})^T)$.

Since $\mathbf{u}$ is divergence free, u_2 has a trace on the interface whose $H^{-1/2}$-norm can be bounded by $\|\mathbf{u}\|_{L^2(\Omega_1)}$ [15], hence the last term on the right side can be bounded by $\|f_4\|_{H^{1/2}(\Gamma)}\|\mathbf{u}\|_{L^2(\Omega_1)}$. The third term is bounded by

$$
\|f_3\|_{L^2(\Gamma)}\|\mathbf{u}\|_{L^2(\Gamma)} \leqq \|f_3\|_{L^2(\Gamma)}\|\mathbf{u}\|_{L^2(\Omega_1)}^{1/2}\|\mathbf{u}\|_{H^1(\Omega_1)}^{1/2}.
$$

We assume that the right side of (3.9) is bounded by a constant of order one, and we wish to bound the left side. From (3.10), we immediately obtain bounds for $\|\mathbf{u}\|_{H^1}$, $\|\mathbf{v}\|_{H^1}$, $\|h\|_{H^1}$ (we have used Korn's inequality here).

In the following, we make repeated use of the following estimates

$$
\begin{aligned}
(3.11)\qquad & \|\mathbf{u}\|_{H^2} + \|\mathbf{v}\|_{H^2} + \|p\|_{H^1} + \|q\|_{H^1} \\
& \quad \leqq C\left(\|\mathbf{f}_1\|_{L^2} + \|\mathbf{f}_2\|_{L^2} + \|f_3\|_{H^{1/2}} + |\lambda|\|h\|_{H^{3/2}} + |\lambda|\|\mathbf{u}\|_{L^2} + |\lambda|\|\mathbf{v}\|_{L^2}\right),
\end{aligned}
$$

$$
(3.12)\qquad \|h\|_{h^{5/2}} \leqq C\left(\|\mathbf{u}\|_{H^2} + \|\mathbf{v}\|_{H^2} + \|p\|_{H^1} + \|q\|_{H^1} + \|f_4\|_{H^{1/2}}\right),
$$

$$
(3.13)\qquad \|h\|_{H^{3/2}} \leqq \|h\|_{H^1}^{2/3}\|h\|_{H^{5/2}}^{1/3}.
$$

Estimate (3.11) follows from (3.1)–(3.6) and (3.8), which form an elliptic system in the sense of Agmon, Douglis and Nirenberg [1]. This can be seen as follows: We can formally regard (3.1), (3.2) as being posed in the same domain by mapping the strip occupied by fluid 2 onto the strip occupied by fluid 1. We do this in such a way that the interface is mapped onto itself and the solid boundaries are mapped onto each other. This yields a system for the six unknowns $(u_1, u_2, p, v_1, v_2, q)$, which are now defined on the same domain. That the Stokes equations are an elliptic system is well known. The same holds of course for two sets of Stokes equations. It is also well known that Dirichlet boundary conditions satisfy the complementing condition. Showing the complementing nature of the interface condition is a straightforward calculation, which we omit. Equation (3.12) is a trivial consequence of (3.7) and the trace theorem, and (3.13) follows from the convexity property of Sobolev norms.

In the following, we start from the energy equation and then use (3.11)–(3.13) in an iterative fashion to obtain better and better estimates. This will lead to the following preliminary result.

Lemma 3.2. *For any* $\varepsilon > 0$, *the following quantities can be estimated by the right-hand side of* (3.9) (*with constants allowed to depend on* ε):

$$\|h\|_{H^1}|\lambda|^{1-\varepsilon},\quad \|\mathbf{u}\|_{L^2}|\lambda|^{1-\varepsilon},\quad \|\mathbf{v}\|_{L^2}|\lambda|^{1-\varepsilon},\quad \|\mathbf{u}\|_{H^1}|\lambda|^{1/2-\varepsilon},\quad \|\mathbf{v}\|_{H^1}|\lambda|^{1/2-\varepsilon},$$
$$\left(\|\mathbf{u}\|_{H^2}+\|p\|_{H^1}\right)|\lambda|^{-1/2-\varepsilon},\quad \left(\|\mathbf{v}\|_{H^2}+\|q\|_{H^1}\right)|\lambda|^{-1/2-\varepsilon}.$$

Proof. We prove by induction that

$$(3.14)_n \qquad \|h\|_{H^1} \leqq C|\lambda|^{2^n/3^n-1},$$

$$(3.15)_n \qquad \|\mathbf{u}\|_{L^2},\ \|\mathbf{v}\|_{L^2} \leqq C|\lambda|^{2^n/3^n-1},$$

$$(3.16)_n \qquad \|\mathbf{u}\|_{H^1},\ \|\mathbf{v}\|_{H^1} \leqq C|\lambda|^{1/2(2^n/3^n-1)},$$

$$(3.17)_n \qquad \|\mathbf{u}\|_{H^2}+\|p\|_{H^1},\ \|\mathbf{v}\|_{H^2}+\|q\|_{H^1} \leqq C|\lambda|^{1/2+2^n/3^n}.$$

Obviously the lemma follows by letting $n \to \infty$. We already have (3.14)–(3.16) for $n=0$. By combining (3.11)–(3.13), we obtain

$$(3.18)\quad \|\mathbf{u}\|_{H^2}+\|p\|_{H^1}+\|\mathbf{v}\|_{H^2}+\|q\|_{H^1}$$
$$\leqq C\Big[\|\mathbf{f}_1\|_{L^2}+\|\mathbf{f}_2\|_{L^2}+\|f_3\|_{H^{1/2}}+|\lambda|\|\mathbf{u}\|_{L^2}+|\lambda|\|\mathbf{v}\|_{L^2}$$
$$+|\lambda|\|h\|_{H^1}^{2/3}\cdot\left(\|\mathbf{u}\|_{H^2}+\|\mathbf{v}\|_{H^2}+\|p\|_{H^1}+\|q\|_{H^1}+\|f_4\|_{H^{1/2}}\right)^{1/3}\Big].$$

With $\beta=\|\mathbf{u}\|_{H^2}+\|p\|_{H^1}+\|\mathbf{v}\|_{H^2}+\|q\|_{H^1}$, we find, using $(3.14)_n$–$(3.16)_n$

$$(3.19)\qquad \beta \leqq C\left(1+|\lambda|^{2^n/3^n}+|\lambda|^{2^{n+1}/3^{n+1}+1/3}\left(\beta+c'\right)^{1/3}\right).$$

From this, $(3.17)_n$ follows easily.

Next, we wish to show that $(3.14)_{n+1}$–$(3.16)_{n+1}$ follow from $(3.17)_n$. Using (3.8), the trace theorem and the convexity property of Sobolev norms, we find

$$(3.20)\qquad \|h\|_{H^1} \leqq \frac{C}{|\lambda|}\sqrt{\|\mathbf{u}\|}_{H^2}\sqrt{\|\mathbf{u}\|}_{H^1}.$$

Moreover (3.10) implies

$$(3.21)\qquad \|\mathbf{v}\|_{H^1}+\|\mathbf{u}\|_{H^1} \leqq C\left(\|\mathbf{u}\|_{L^2}+\|\mathbf{v}\|_{L^2}+|\lambda|^{-1/4}\|\mathbf{u}\|_{L^2}^{1/2}\|\mathbf{u}\|_{H^1}^{1/2}\right)^{1/2},$$

and

$$(3.22)$$
$$\|\mathbf{v}\|_{L^2}+\|\mathbf{u}\|_{L^2} \leqq C\Big[\|h\|_{H^1}+|\lambda|^{-1/2}\left(\|\mathbf{u}\|_{L^2}+\|\mathbf{v}\|_{L^2}+|\lambda|^{-1/4}\|\mathbf{u}\|_{L^2}^{1/2}\|\mathbf{u}\|_{H^1}^{1/2}\right)^{1/2}\Big].$$

With $x=\|\mathbf{u}\|_{H^1}+\|\mathbf{v}\|_{H^1}$, $y=\|\mathbf{u}\|_{L^2}+\|\mathbf{v}\|_{L^2}$ and using $(3.17)_n$, we find that (3.21), (3.22) take the following form

$$(3.23)\qquad x \leqq C\left(y^{1/2}+|\lambda|^{-1/8}x^{1/4}y^{1/4}\right),$$

$$(3.24)\qquad y \leqq C\left(|\lambda|^{2^{n-1}/3^n-3/4}x^{1/2}+|\lambda|^{-1/2}y^{1/2}+|\lambda|^{-5/8}y^{1/4}x^{1/4}\right).$$

From (3.23) it follows that $x \leqq Cy^{1/2}$ or $x \leqq C|\lambda|^{-1/6}y^{1/3}$, depending on whether the first or second term on the right is bigger. In the first case, (3.24) yields

$$y \leqq C\left(|\lambda|^{2^{n-1}/3^n - 3/4} y^{1/4} + |\lambda|^{-1/2} y^{1/2} + |\lambda|^{-5/8} y^{3/8}\right). \tag{3.25}$$

From this $(3.15)_{n+1}$ is immediate. In the second case, we get

$$y \leqq C\left(|\lambda|^{2^{n-1}/3^n - 3/4 - 1/12} y^{1/6} + |\lambda|^{-1/2} y^{1/2} + |\lambda|^{-5/8 - 1/24} y^{1/3}\right). \tag{3.26}$$

From this, $(3.15)_{n+1}$ is also immediate. From $(3.15)_{n+1}$ and (3.21) follows $(3.16)_{n+1}$, and using (3.20) and $(3.17)_n$ we find $(3.14)_{n+1}$. This concludes the proof of the lemma.

We now return to the proof of Theorem 3.1.

To proceed further, we take difference quotients in the x-direction. These satisfy the same equations (3.1)–(3.8) with the f's replaced by their difference quotients. From (3.10), we then see that the H^1-norms of all x-derivatives of $\mathbf{u}$ and $\mathbf{v}$ can be estimated by terms of order 1. The divergence condition now yields

$$\|u_2\|_2, \|v_2\|_2 \leqq C.$$

Equation (3.8) and the trace theorem imply that $\|h\|_{H^{3/2}} \leqq C/|\lambda|$, and from (3.11), and Lemma 3.2 we conclude that

$$\|\mathbf{u}\|_{H^2} + \|\mathbf{v}\|_{H^2} + \|p\|_{H^1} + \|q\|_{H^1} \leqq C|\lambda|^{\varepsilon}.$$

Using this in (3.20), we get $\|h\|_{H^1} \leqq C|\lambda|^{-5/4+\varepsilon}$, and by inserting this in (3.21), (3.22), we find $\|\mathbf{u}\|_{L^2} + \|\mathbf{v}\|_{L^2} \leqq C/|\lambda|$. By using (3.11) again, we obtain the theorem.

Remarks. We have so far only given estimates for a solution that was assumed to exist and have the regularity implied by the left-hand side of (3.9). Such estimates show that (3.1)–(3.8) for $\mathrm{Re}\lambda \geqq \gamma > 0$ is solvable for a closed set of f's (in the topology indicated by the right side of (3.9)). Solvability for a dense set of f's can be shown in a straightforward manner by separation of variables. (Separation of variables leads to a system of ODE's, and it is easy to show that a Fredholm alternative holds for these ODE's. The absence of eigenfunctions follows from the energy equation.) From this we see that in fact, for any λ with $\mathrm{Re}\lambda > 0$, (3.1)–(3.8) has a unique solution. Inequality (3.9) holds uniformly in any closed subset of any right half-plane, if this subset contains no eigenvalues. Moreover, by compactness, the number of eigenvalues is countable, and there can be only finitely many in any bounded set. Equation (3.10) implies that all eigenvalues have negative real parts.

4. Bifurcation to travelling waves. It is convenient to use a domain mapping which takes the domain occupied by each fluid to a fixed one. The most straightforward way to construct such a mapping is to stretch or contract vertical lines. We shall in addition transform the velocity fields in such a way that the divergence condition is preserved and (2.8) reduces to the linearized form even in the nonlinear case. In doing this, we essentially follow Beale [2].

Let $y = h_0$ be the flat interface of the rest state, and let $y = h(x,t)$ be the actual interface, which we assume periodic in x with period L.

Let P be any linear extension operator that maps functions $h(\zeta) \in H^s(\Gamma)$, into functions $\tilde{h}(\zeta,\eta)$ such that $\tilde{h}(\zeta,h_0) = h(\zeta)$ and $\tilde{h} \in H^{s+1/2}(\Omega)$ (P exist according to the trace theorem). For simplicity, we also assume that P takes $h \equiv h_0$ to $\tilde{h} \equiv h_0$. Let then $f_0(\eta)$ be a C^∞-function of η such that $f_0 = 1$ near $\eta = h_0$ and $f_0 = 0$ near $\eta = 0$ and $\eta = 1$. Define $\bar{h}(\zeta,\eta,t) = \tilde{h}(\zeta,\eta,t)\cdot f_0(\eta) + h_0(1 - f_0(\eta))$.

We now define

$$\theta(\zeta,\eta,t)=\left(\zeta,\eta\cdot\frac{\bar{h}(\zeta,\eta,t)}{h_0}\right). \tag{4.1}$$

Evidently, θ maps the strip $0\leqq\eta\leqq h_0$ to Ω_1, and the strip $h_0\leqq\eta\leqq 1$ to Ω_2. The velocities are transformed as follows:

$$u_i(\theta(\zeta,\eta,t))=\frac{\partial\theta_i}{\partial\zeta_j}\tilde{u}_j/J, \tag{4.2}$$

where J is the Jacobian of θ. Of course, $\mathbf{v}$ in Ω_2 is transformed in the same way. Explicitly, (4.2) reads

$$\mathbf{u}=\left(\frac{\bar{h}}{h_0}+\eta\frac{\bar{h}_\eta}{h_0}\right)^{-1}\left(\tilde{u}_1,\eta\frac{\bar{h}_\zeta}{h_0}\tilde{u}_1+\left(\frac{\bar{h}}{h_0}+\eta\frac{\bar{h}_\eta}{h_0}\right)\tilde{u}_2\right). \tag{4.3}$$

Formula (4.2) is set up in such a way that $\mathbf{u}$, $\mathbf{v}$ are divergence-free in the x, y-plane, if $\tilde{\mathbf{u}}$, $\tilde{\mathbf{v}}$ are divergence-free in the ζ, η-plane. Moreover, (2.8) now assumes the simple form

$$\dot{h}(\zeta,t)=\tilde{u}_2(\zeta,h_0,t). \tag{4.4}$$

It is also clear that (2.5) does not change its form, i.e. $\mathbf{u}=\mathbf{v}$ at $y=h(x)$ simply becomes $\tilde{\mathbf{u}}=\tilde{\mathbf{v}}$ at $\eta=h_0$. The boundary conditions at the walls are also preserved. When these substitutions are inserted into (2.1)–(2.8), we obtain a new set of equations, which we do not write down explicitly. We shall refer to these new equations as (2.1)*–(2.8)*. If we have a flat interface $h=h_0$, then of course (4.3) reads $\mathbf{u}=\tilde{\mathbf{u}}$, and (2.1)*–(2.8)* have the same form as (2.1)–(2.8).

Plane Couette flow is the following solution of (2.1)–(2.8):

$$\hat{h}(x)\equiv h_0,\ \hat{u}_1=\frac{\eta_2 V_0}{\eta_1+h_0(\eta_2-\eta_1)}y,\quad \hat{u}_2=0,\quad \hat{v}_1=\frac{\eta_1 V_0 y+V_0 h_0(\eta_2-\eta_1)}{\eta_1+h_0(\eta_2-\eta_1)},$$
$$\hat{v}_2=0,\quad \hat{p}=0,\quad \hat{q}=0.$$

We can linearize at this solution, and obtain a set of linearized equations analogous to (3.1)–(3.8). As usual, we call λ an eigenvalue, if the homogeneous linearized problem has nontrivial solutions. The estimates in §3 imply that, for $V_0=0$ (rest), there is a countable sequence of eigenvalues, all in a sector of the left half plane bounded away from the imaginary axis. All these eigenvalues have finite multiplicity and Fredholm index zero. Estimates like those in §3 can easily be extended to the linearization at Couette flow with finite V_0. If we linearize (2.1)–(2.8) at this flow, then there are a number of terms perturbing (3.1)–(3.8). All these terms are relatively compact except the one resulting from u_1h' in (2.8). This latter term vanishes in a frame moving with the fluid on the interface. Standard perturbation theory [12] can now be used to show that estimates like in §3 hold for λ in any closed set that lies in a right half plane and contains no eigenvalues. However, there can, and as [6], [16] show, there will be a finite number of eigenvalues with positive real parts if V_0 is large enough. Generically, there will be a critical value V_{0c}, such that, for $V_0<V_{0c}$, all eigenvalues have negative real parts, but a pair of simple complex conjugate eigenvalues crosses the imaginary axis transversally as V_0 increases past V_{0c}. Let us denote these imaginary eigenvalues by $\pm i\omega_0$.

We introduce the following substitutions in (2.1)*–(2.8)*:

$$\tau=\omega t,\quad \mathbf{u}^*=\tilde{\mathbf{u}}-\hat{\mathbf{u}}(V_0),\quad \mathbf{v}^*=\tilde{\mathbf{v}}-\hat{\mathbf{v}}(V_0),\quad h^*=(h-h_0)\frac{\omega}{\omega_0}.$$

We use the following notation for function space: $H^k_{\Omega_1}$ denotes the space of functions defined on $-\infty<\zeta<\infty$, $0\leqq\eta\leqq h_0$, which have period L in ζ and H^k-regularity. Similarly, we define $H^k_{\Omega_2}$. Finally H^k_Γ is the set of k times differentiable periodic functions depending on ζ alone.

$H^k(X)$ denotes the spaces of all 2π-periodic functions defined on $-\infty<\tau<\infty$, taking values in X, and having k square integrable derivatives.

For the analysis of (2.1)*–(2.8)*, we choose the following space V:

$$\begin{aligned}V=\Big\{&(\mathbf{u}^*,\mathbf{v}^*,p,q,h^*)\,|\,\mathbf{u}^*\in H^1\big(H^2_{\Omega_1}\big)\cap H^2\big(L^2_{\Omega_1}\big),\ v^*\in H^1\big(H^2_{\Omega_2}\big)\cap H^2\big(L^2_{\Omega_2}\big),\\ &p\in H^1\big(H^1_{\Omega_1}\big),\ q\in H^1\big(H^1_{\Omega_2}\big),\ h^*\in H^1\big(H^{5/2}_\Gamma\big)\cap H^2\big(H^{3/2}_\Gamma\big);\\ &\operatorname{div}\mathbf{u}^*=0,\ \operatorname{div}\mathbf{v}^*=0,\\ &\iint_{\Omega_1}p+\iint_{\Omega_2}q=0,\ \mathbf{u}^*=0 \text{ at } \eta=0,\ \mathbf{v}^*=0 \text{ at } \eta=1,\ \mathbf{u}^*=\mathbf{v}^* \text{ at}\\ &\qquad\eta=h_0,\ \omega_0\frac{\partial h^*}{\partial\tau}=\mathbf{u}^* \text{ at } \eta=h_0,\ \int_0^L h^*(\zeta)\,d\zeta=0\Big\}.\end{aligned}$$

Functions in this space have sufficient regularity such that all the nonlinearities in (2.1)*–(2.8)* are defined. We can now prove a Hopf bifurcation result based on the implicit function theorem. This relies essentially on an iterative scheme which at each stage solves the linearized problem with the nonlinear terms as inhomogeneities. It is important that such an iteration takes the space V into itself, i.e. that by inverting the linearized operator we gain at least as much regularity as we lose by evaluating the nonlinear terms. This is guaranteed by the coercive estimate of Theorem 3.1. In this way, we obtain the following.

THEOREM 4.1. *Assume that, at $V_0=V_{0c}$, there is a pair of algebraically simple complex conjugate eigenvalues $\pm i\omega_0$, $\omega_0\neq 0$, and no other eigenvalue is an integral multiple of $i\omega_0$. Moreover, assume that those eigenvalues cross the imaginary axis transversally, i.e. if $\lambda(V_0)$ denotes the branch of eigenvalues for which $\lambda(V_{0c})=i\omega_0$, then $(d/dV_0)\operatorname{Re}\lambda(V_0)\,|_{V_0=V_{0c}}\neq 0$. Then there is an analytic branch of nontrivial time-periodic solutions $(\mathbf{u}^*(\varepsilon),\mathbf{v}^*(\varepsilon),p(\varepsilon),q(\varepsilon),h^*(\varepsilon))=:Y(\varepsilon)$, such that $Y(\varepsilon)\in V$ is a solution of (2.1)*–(2.8)* for $V_0=V_0(\varepsilon)$ with temporal frequency $\omega=\omega(\varepsilon)$. For $\varepsilon=0$, we have $V_0=V_{0c}$, $\omega=\omega_0$ and $Y=0$. This branch of periodic solutions is unique except for phase shift or changes of parametrization.*

If, at $V_0=V_{0c}$, all eigenvalues other than $\pm i\omega_0$ have negative real parts and we have $\frac{d}{dV_0}\operatorname{Re}\lambda(V_0)\,|_{V_0=V_{0c}}>0$, then the bifurcating periodic solutions are stable if $V_0(\varepsilon)>V_{0c}$ for small $\varepsilon\neq 0$, and unstable if $V_0(\varepsilon)<V_{0c}$.

Remark. It is easy to show higher regularity of the bifurcating solutions by choosing function spaces of higher regularity for the bifurcation analysis.

5. Reduction of the bifurcation problem to local form. In the previous two sections, we have provided the analytical tools and the estimates needed to establish a bifurcation theorem. In the following, we now describe an algorithm for calculating approximations to this bifurcating solution.

As before, we study bifurcation from plane Couette flow, and we consider the velocity of the upper plate as the bifurcation parameter. Plane Couette flow is the following solution of (2.1)–(2.8):

$$
\begin{aligned}
&\hat{h}(x)=h_0,\\
&\hat{u}_1=\frac{\eta_2 V_0 y}{\eta_1+h_0(\eta_2-\eta_1)},\\
&\hat{u}_2=0, \qquad \hat{p}=0,\\
&\hat{v}_1=\frac{[\eta_1 y+h_0(\eta_2-\eta_1)]V_0}{\eta_1+h_0(\eta_2-\eta_1)},\\
&v_2=0, \qquad q=0.
\end{aligned}
\tag{5.1}
$$

For the bifurcation problem, it is convenient to introduce new variables representing the perturbation of this solution. We therefore replace u_1 and v_1 by $u_1+\hat{u}_1$, $v_1+\hat{v}_1$, where $\hat{u}_1$, $\hat{v}_1$ are given by formula (5.1) in the regions $0\leqq y\leqq h(x,t)$ and $h(x,t)\leqq y\leqq 1$, respectively. Moreover, we set

$$
h(x,t)=h_0+\delta(x,t), \tag{5.2}
$$

and $\delta(x,t)$ has zero mean value as a function of x. With this change of variables, (2.1) and (2.2) take the form

$$
\rho\left[\dot{\mathbf{u}}+\hat{u}_1\frac{\partial \mathbf{u}}{\partial x}+\mathbf{e}_x u_2\frac{\partial \hat{u}_1}{\partial y}+(\mathbf{u}\cdot\nabla)\mathbf{u}\right]=\eta_1\Delta\mathbf{u}-\nabla p, \quad 0\leqq y\leqq h(x,t), \quad \operatorname{div}\mathbf{u}=0, \tag{5.3}
$$

$$
\rho\left[\dot{\mathbf{v}}+\hat{v}_1\frac{\partial \mathbf{v}}{\partial x}+\mathbf{e}_x v_2\frac{\partial \hat{v}_1}{\partial y}+(\mathbf{v}\cdot\nabla)\mathbf{v}\right]=\eta_2\Delta\mathbf{v}-\nabla q, \quad h(x,t)\leqq y\leqq 1, \quad \operatorname{div}\mathbf{v}=0. \tag{5.4}
$$

On the walls, we have

$$
\mathbf{u}=0 \text{ at } y=0, \qquad \mathbf{v}=0 \text{ at } y=1. \tag{5.5}
$$

In the normal and shear stress conditions, the terms replaced by $\hat{u}_1$, $\hat{v}_1$ are such that they cancel. Moreover, $h'=\delta'$, so h can be replaced by δ. Across the interface $y=h(x,t)$, we thus obtain the following conditions resulting from (2.6) and (2.7)

$$
\begin{aligned}
&(1-\delta'^2)\eta_1\left(\frac{\partial u_2}{\partial x}+\frac{\partial u_1}{\partial y}\right)+2\eta_1\delta'\left(\frac{\partial u_2}{\partial y}-\frac{\partial u_1}{\partial x}\right)\\
&\qquad=(1-\delta'^2)\eta_2\left(\frac{\partial v_2}{\partial x}+\frac{\partial v_1}{\partial y}\right)+2\eta_2\delta'\left(\frac{\partial v_2}{\partial y}-\frac{\partial v_1}{\partial x}\right),
\end{aligned}
\tag{5.6}
$$

$$
\begin{aligned}
&2\eta_1\frac{\partial u_2}{\partial y}-p-2\delta'\eta_1\left(\frac{\partial u_2}{\partial x}+\frac{\partial u_1}{\partial y}\right)+\delta'^2\left(2\eta_1\frac{\partial u_1}{\partial x}-p\right)\\
&\qquad=2\eta_2\frac{\partial v_2}{\partial y}-q-2\delta'\eta_2\left(\frac{\partial v_2}{\partial x}+\frac{\partial v_1}{\partial y}\right)+\delta'^2\left(2\eta_2\frac{\partial v_1}{\partial x}-q\right)+T\delta''/(1+\delta'^2)^{1/2}.
\end{aligned}
\tag{5.7}
$$

The condition (2.5) is replaced by

$$u_1 - v_1 = \hat{v}_1(h_0+\delta) - \hat{u}_1(h_0+\delta) = \frac{(\eta_1-\eta_2)\delta V_0}{\eta_1 + h_0(\eta_2-\eta_1)}, \tag{5.8}$$

$$u_2 = v_2. \tag{5.9}$$

Equation (5.8) shows that, for $\eta_1 \neq \eta_2$ and $V_0 \neq 0$, $\delta(x,t)$ can be eliminated and expressed by the jump in the x-component of velocity at $y = h_0 + \delta(x,t)$. Finally, equation (2.8) assumes the form

$$\dot{\delta} + \hat{u}_1(h_0)\delta' + u_1\delta' + \frac{\eta_2 V_0 \delta\delta'}{\eta_1 + h_0(\eta_2-\eta_1)} = u_2. \tag{5.10}$$

Our bifurcation problem is now given by (5.1)–(5.10). The null solution corresponds to plane Couette flow. We consider δ in (5.6), (5.7) and (5.10) as having been eliminated from (5.8), and look for solutions $(\mathbf{u}, \mathbf{v}, p, q)$ which are periodic in x.

6. The spectral problem and its adjoint. The spectral problem for the stability of the null solution is

$$\rho\left[\lambda\mathbf{u} + \hat{u}_1 \frac{\partial \mathbf{u}}{\partial x} + \mathbf{e}_x u_2 \hat{u}_1'\right] + \nabla p - 2\eta_1 \operatorname{div} \underline{\underline{D}}[\mathbf{u}] = 0, \quad 0 \leqq y \leqq h_0, \quad \operatorname{div}\mathbf{u} = 0, \tag{6.1}$$

$$\rho\left[\lambda\mathbf{v} + \hat{v}_1 \frac{\partial \mathbf{v}}{\partial x} + \mathbf{e}_x v_2 \hat{v}_1'\right] + \nabla q - 2\eta_2 \operatorname{div} \underline{\underline{D}}[\mathbf{v}] = 0, \quad h_0 \leqq y \leqq 1, \quad \operatorname{div}\mathbf{v} = 0. \tag{6.2}$$

Here $\underline{\underline{D}}[\mathbf{u}]$ is the symmetric part of $\nabla\mathbf{u}$. $\hat{u}_1'$ and $\hat{v}_1'$ denote derivatives with respect to h, while h' and δ' will continue to denote derivatives with respect to x. On the walls we have

$$\mathbf{u} = 0 \text{ at } y = 0, \qquad \mathbf{v} = 0 \text{ at } y = 1. \tag{6.3}$$

From (5.8), we have $\delta = k(u_1 - v_1)$, where $k = (\eta_1 + (\eta_2 - \eta_1)h_0)/V_0(\eta_1 - \eta_2)$.

By inserting this into the remaining interface conditions and linearizing, we find the following conditions at $y = h_0$:

$$u_2 - v_2 = 0, \tag{6.4}$$

$$2\eta_1 D_{12}[\mathbf{u}] - 2\eta_2 D_{12}[\mathbf{v}] = 0, \tag{6.5}$$

$$-p + 2\eta_1 D_{22}[\mathbf{u}] + q - 2\eta_2 D_{22}[\mathbf{v}] - Tk(u_1'' - v_1'') = 0, \tag{6.6}$$

$$\left(\lambda + \hat{u}(h_0)\frac{\partial}{\partial x}\right) k(u_1 - v_1) - u_2 = 0. \tag{6.7}$$

We turn next to the computation of the spectral problem, which is adjoint to (6.1)–(6.7). We multiply (6.1) by $\bar{\mathbf{u}}^*$, (6.2) by $\bar{\mathbf{v}}^*$, the complex conjugates of the adjoint velocities, and integrate the resulting expressions over their domain of definition. We assume periodicity in x with period L, and integrate by parts using periodicity, solenoidality

and (6.3) to derive

$$(6.8)\qquad \int_{\Omega_1}\left\{\rho\lambda\bar{\mathbf{u}}^*-\rho\hat{u}_1\frac{\partial\bar{\mathbf{u}}^*}{\partial x}+\rho\bar{u}_1^*\hat{u}_1'\mathbf{e}_y-2\eta_1\operatorname{div}\underline{\underline{D}}[\bar{\mathbf{u}}^*]\right\}\cdot\mathbf{u}\,dx\,dy$$

$$+\int_{\Omega_2}\left\{\rho\lambda\bar{\mathbf{v}}^*-\rho\hat{v}_1\frac{\partial\bar{\mathbf{v}}^*}{\partial x}+\rho\bar{v}_1^*\hat{v}_1'\mathbf{e}_y-2\eta_2\operatorname{div}\underline{\underline{D}}[\bar{\mathbf{v}}^*]\right\}\cdot\mathbf{v}\,dx\,dy$$

$$-\int_{\Omega_1}p\operatorname{div}\bar{\mathbf{u}}^*-\int_{\Omega_2}q\operatorname{div}\bar{\mathbf{v}}^*$$

$$=-\int_0^L\{(-q+2\eta_2D_{22}[\mathbf{v}])\bar{v}_2^*-(-p+2\eta_1D_{22}[\mathbf{u}])\bar{u}_2^*$$

$$+2\eta_2\bar{v}_1^*D_{12}[\mathbf{v}]-2\eta_2v_1D_{12}[\bar{\mathbf{v}}^*]-2\eta_2v_2D_{22}[\bar{\mathbf{v}}^*]$$

$$+2\eta_1u_1D_{12}[\bar{\mathbf{u}}^*]-2\eta_1\bar{u}_1^*D_{12}[\mathbf{u}]+2\eta_1u_2D_{22}[\mathbf{u}^*]\}\,dx.$$

By considering special forms of $\mathbf{u}$, $\mathbf{v}$, p, q, we find that in $0\leqq y\leqq h_0$, we have

$$(6.9)\qquad \rho\lambda\bar{\mathbf{u}}^*-\rho\hat{u}_1\frac{\partial\bar{\mathbf{u}}^*}{\partial x}+\rho\mathbf{e}_y\bar{u}_1^*\hat{u}_1'-2\eta_1\operatorname{div}\underline{\underline{D}}[\bar{\mathbf{u}}^*]=-\nabla\bar{p}^*,\qquad \operatorname{div}\bar{\mathbf{u}}^*=0,$$

whilst in $h_0\leqq y\leqq 1$,

$$(6.10)\qquad \rho\lambda\bar{\mathbf{v}}^*-\rho\hat{v}_1\frac{\partial\bar{\mathbf{v}}^*}{\partial x}+\rho\mathbf{e}_y\bar{v}_1^*\hat{v}_1'-2\eta_2\operatorname{div}\underline{\underline{D}}[\bar{\mathbf{v}}^*]=-\nabla q^*,\qquad \operatorname{div}\bar{\mathbf{v}}^*=0.$$

We insert (6.9) and (6.10) back into (6.8), and compute

$$\int_{\Omega_1}\mathbf{u}\cdot\nabla\bar{p}^*\,dx\,dy+\int_{\Omega_2}\mathbf{v}\cdot\nabla\bar{q}^*\,dx\,dy=\int_0^L(\bar{p}^*u_2-\bar{q}^*v_2)\,dx.$$

This term is added to the right-hand side of (6.8), leading to

$$(6.11)\qquad 0=\int_0^L\{(-q+2\eta_2D_{22}[\mathbf{v}])\bar{v}_2^*+(p-2\eta_1D_{22}[\mathbf{u}])\bar{u}_2^*$$

$$+(\bar{q}^*-2\eta_2D_{22}[\bar{\mathbf{v}}^*])v_2+(-\bar{p}^*+2\eta_1D_{22}[\bar{\mathbf{u}}^*])u_2$$

$$+2\eta_2\bar{v}_1^*D_{12}[\mathbf{v}]-2\eta_1\bar{u}_1^*D_{12}[\mathbf{u}]$$

$$+2\eta_1u_1D_{12}[\bar{\mathbf{u}}^*]-2\eta_2v_1D_{12}[\bar{\mathbf{v}}^*]\}\,dx.$$

We use (6.6) to reduce the first line of (6.11), (6.4) for the second line and (6.5) for the third line. Thus we find

$$(6.12)\quad 0=\int_0^L\{(-q+2\eta_2D_{22}[\mathbf{v}])(\bar{v}_2^*-\bar{u}_2^*)-T\delta''\bar{u}_2^*$$

$$+u_2(\bar{q}^*-2\eta_2D_{22}[\bar{\mathbf{v}}^*]-\bar{p}^*+2\eta_1D_{22}[\bar{\mathbf{u}}^*])$$

$$+(\bar{v}_1^*-\bar{u}_1^*)2\eta_1D_{12}[\mathbf{u}]$$

$$+(u_1-v_1)\cdot2\eta_1D_{12}[\bar{\mathbf{u}}^*]+v_1(2\eta_1D_{12}[\bar{\mathbf{u}}^*]-2\eta_2D_{12}[\bar{\mathbf{v}}^*])\}\,dx.$$

We next write

$$\int_0^L \delta'' \bar{u}_2^* \, dx = \int_0^L \delta \bar{u}_2^{*\prime\prime} \, dx,$$

and set $\delta = k(u_1 - v_1)$, $u_2 = (\lambda + \hat{u}(h_0)\partial/\partial x)k(u_1 - v_1)$.

This leads to

$$\begin{aligned} 0 = \int_0^L \Big\{ & \big(-q + 2\eta_2 D_{22}[\mathbf{v}]\big)\big(\bar{v}_2^* - \bar{u}_2^*\big) \\ & + (u_1 - v_1)\Big\langle -Tk\bar{u}_2^{*\prime\prime} + \Big(\lambda - \hat{u}(h_0)\frac{\partial}{\partial x}\Big)k\big(\bar{q}^* - 2\eta_2 D_{22}[\bar{\mathbf{v}}^*] \\ & \qquad - \bar{p}^* + 2\eta_1 D_{22}[\bar{\mathbf{u}}^*]\big) + 2\eta_1 D_{12}[\bar{\mathbf{u}}^*]\Big\rangle \\ & + \big(\bar{v}_1^* - \bar{u}_1^*\big)\cdot 2\eta_1 D_{12}[\mathbf{u}] + v_1\big(2\eta_1 D_{12}[\bar{\mathbf{u}}^*] - 2\eta_2 D_{12}[\bar{\mathbf{v}}^*]\big)\Big\} \, dx. \end{aligned}$$

This yields the following four conditions on $y = h_0$:

$$2\eta_1 D_{12}[\bar{\mathbf{u}}^*] = 2\eta_2 D_{12}[\bar{\mathbf{v}}^*], \tag{6.13}$$

$$\bar{u}_1^* = \bar{v}_1^*, \tag{6.14}$$

$$\bar{u}_2^* = \bar{v}_2^*, \tag{6.15}$$

$$\begin{aligned} -Tk\bar{u}_2^{*\prime\prime} + \Big(\lambda - \hat{u}(h_0)\frac{\partial}{\partial x}\Big)k\big(\bar{q}^* - 2\eta_2 D_{22}[\bar{\mathbf{v}}^*] \\ - \bar{p}^* + 2\eta_1 D_{22}[\bar{\mathbf{u}}^*]\big) + 2\eta_1 D_{12}[\bar{\mathbf{u}}^*] = 0. \end{aligned} \tag{6.16}$$

Thus the adjoint problem is given by the differential equations (6.9), (6.10), the Dirichlet conditions $\bar{\mathbf{u}}^* = 0$, $\bar{\mathbf{v}}^* = 0$ on the walls and conditions (6.13)–(6.16) on the interface.

It is easy to establish a necessary and sufficient condition for the solvability of the inhomogeneous problem corresponding to (6.1)–(6.7). Suppose that the zeros on the right of the first equation in (6.1) and (6.2) and on the right of (6.4)–(6.7) are replaced by

$$\mathbf{g}_1(x,y), \quad \mathbf{g}_2(x,y), \quad g_3(x), \quad g_4(x), \quad g_5(x), \quad g_6(x), \tag{6.17}$$

respectively. This inhomogeneous problem is solvable if and only if the data (6.17) are orthogonal to the kernel of the adjoint, that is, when (6.17) is such that

$$\begin{aligned} & \int_{\Omega_1} \mathbf{g}_1 \cdot \bar{\mathbf{u}}^* \, dx\, dy + \int_{\Omega_2} \mathbf{g}_2 \cdot \bar{\mathbf{v}}^* \, dx\, dy \\ & \quad = \int_0^L \big\{ g_3\big(\bar{q}^* - 2\eta_2 D_{22}[\bar{\mathbf{v}}^*]\big) + g_4 \bar{u}_1^* + g_5 \bar{u}_2^* \\ & \qquad + g_6\big(\bar{q}^* - 2\eta_2 D_{22}[\bar{\mathbf{v}}^*] - \bar{p}^* + 2\eta_1 D_{22}[\bar{\mathbf{u}}^*]\big)\big\} \, dx. \end{aligned} \tag{6.18}$$

We are interested in the neighborhood of a critical point, where a loss of stability occurs. There is a critical plate velocity $V_0 = \hat{V}_0$ such that the real part of λ vanishes, and $\lambda = i\omega_0$. We put

$$V_0 = \hat{V}_0(1 + R),$$

so criticality is when $R=0$. We get Hopf bifurcation when the loss of stability is strict

$$\operatorname{Re} \left.\frac{\partial \lambda}{\partial R}\right|_{R=0} \neq 0.$$

Let

$$\zeta_0 = \begin{cases} \mathbf{u} & \text{in } \Omega_1 (y < h_0), \\ \mathbf{v} & \text{in } \Omega_2 (y > h_0) \end{cases}$$

be a right eigenvector satisfying (6.1)–(6.7) at criticality. Then $\bar{\zeta}_0$ is the right eigenvector belonging to $-i\omega_0$. $\bar{\zeta}_0^*$ and ζ_0^* are the left eigenvectors belonging to $i\omega_0$ and $-i\omega_0$.

7. Domain perturbations and Hopf bifurcation. We have already demonstrated that Couette flow can bifurcate into a time-periodic solution in which we have travelling interfacial waves. To compute this solution we would, following Lindstedt, map the solution into a fixed frequency domain (2π periodic in s)

$$\omega \, dt = ds,$$

and replace

$$(\dot{\mathbf{u}}, \dot{\mathbf{v}}, \dot{\delta}) \quad \text{with } \omega \left(\frac{\partial \mathbf{u}}{\partial s}, \frac{\partial \mathbf{v}}{\partial s}, \frac{\partial \delta}{\partial s} \right) \tag{7.1}$$

in (5.3), (5.4) and (5.10). We then map our problem into a fixed spatial domain, using a one-to-one linear mapping, which takes boundary points into boundary points

$$y = \left(h(x,t) - h_0\right) \frac{y_0 - 1}{h_0 - 1} + y_0 \begin{cases} h \leqq y \leqq 1, \\ h_0 \leqq y_0 \leqq 1, \end{cases} \tag{7.2}_1$$

and

$$y = \left(1 + \frac{h(x,t) - h_0}{h_0}\right) y_0 \begin{cases} 0 \leqq y \leqq h(x,t), \\ 0 \leqq y_0 \leqq h_0. \end{cases} \tag{7.2}_2$$

We then change variables, putting $x = x_0$ and $y = \tilde{y}(x_0, y_0)$, where $\tilde{y}$ is defined by (7.2) in (5.3)–(5.10). The form of these equations changes under the change of variables. However, following ideas introduced by Joseph [18], [19] we find many simplifications. We shall now explain these simplifications.

First we introduce an amplitude parameter which is conveniently set as a projection

$$\varepsilon = [\mathbf{u}, \mathbf{z}^*],$$

where

$$[\mathbf{a}, \mathbf{b}] \overset{\text{def}}{=} \frac{1}{2\pi} \int_0^{2\pi} \langle \mathbf{a}, \mathbf{b} \rangle \, ds,$$

and $\langle \mathbf{a}, \mathbf{b} \rangle$ are integrals over both regions of the type displayed in (6.8). We are working in the frame of Iooss-Joseph [17, §VIII. 3] and

$$\mathbf{z}^* = e^{is} \zeta_0^*, \tag{7.3}$$

where ζ_0^* is the left eigenvector at criticality which was introduced at the end of the last section.

The bifurcating solution may be computed in a series of powers of ε. Thus

$$\omega(\varepsilon)=\omega_0+\frac{\varepsilon^2}{2}\omega_2+\frac{\varepsilon^4}{4!}\omega_4+\cdots, \tag{7.4}$$

$$V_0(\varepsilon)=\hat{V}_0\left(1+\frac{\varepsilon^2}{2}\hat{V}_2+\frac{\varepsilon^4}{4!}\hat{V}_4+\cdots\right). \tag{7.5}$$

It follows from the classical theory of Hopf bifurcation that ω and V_0 are even functions of ε. We have assumed this in writing (7.4) and (7.5). Moreover,

$$\left.\begin{matrix}\mathbf{u}(x,y,s,\varepsilon)\\ \mathbf{v}(x,y,s,\varepsilon)\\ p(x,y,s,\varepsilon)\\ q(x,y,s,\varepsilon)\end{matrix}\right\}=\sum_{l=0}^{\infty}\frac{\varepsilon^{l+1}}{(l+1)!}\begin{matrix}\mathbf{u}^{[l]}(x_0,y_0,s)\\ \mathbf{v}^{[l]}(x_0,y_0,s)\\ p^{[l]}(x_0,y_0,s)\\ q^{[l]}(x,y,s).\end{matrix} \tag{7.6}$$

The functions of x and y are defined in deformed domains with a wavy interface $h(x,s,\varepsilon)-h_0=\delta(x,s,\varepsilon)$. The functions of x_0 and y_0 were defined in the reference domain with a flat interface at $y_0=h$. The perturbation of the interface $\delta(x,s,\varepsilon)$ can be eliminated from (5.8); that is,

$$\delta(x,s,\varepsilon)=k(u_1-v_1)(x,s,\varepsilon)$$

is an identity for all x, s, ε. The square brackets on the left of (7.6) indicate differentiation following the mapping evaluated at $\varepsilon=0$. For example,

$$\begin{aligned}\mathbf{u}^{[n]}(x_0,y_0,s)&=\left.\frac{d^n}{d\varepsilon^n}\right|_{\varepsilon=0}u(x,y,s,\varepsilon)\\ &=\frac{\partial^n}{\partial\varepsilon^n}u(x_0,y(x_0,y_0,s,\varepsilon),s,\varepsilon).\end{aligned} \tag{7.7}$$

There is a simple differential calculus for domain perturbations. The partial derivatives, holding x_0,y_0 fixed, at $\varepsilon=0$ are also important. For these

$$\mathbf{u}^{\langle n\rangle}(x_0,y_0,s,\varepsilon)=\left[\frac{\partial^n}{\partial\varepsilon^n}u(x_0,y_0,s,\varepsilon)\right]_{\varepsilon=0}. \tag{7.8}$$

The two types of derivatives are related by the chain rule

$$\begin{aligned}&\mathbf{u}^{[1]}=\mathbf{u}^{\langle 1\rangle}+y^{\langle 1\rangle}\frac{\partial\mathbf{u}^{\langle 0\rangle}}{\partial y},\\ &\mathbf{u}^{[2]}=\mathbf{u}^{[2]}+2y^{\langle 1\rangle}\frac{\partial\mathbf{u}^{\langle 1\rangle}}{\partial y}+\left(y^{\langle 1\rangle}\right)^2\frac{\partial^2\mathbf{u}^{\langle 0\rangle}}{\partial y^2}+y^{\langle 2\rangle}\frac{\partial\mathbf{u}^{\langle 0\rangle}}{\partial y},\\ &\mathbf{u}^{[n]}(x_0,y_0,s)=\mathbf{u}^{\langle n\rangle}(x_0,y_0,s)+\text{lower order terms},\end{aligned} \tag{7.9}$$

where

$$y^{\langle n\rangle}(x_0,y_0,s)=\begin{cases}\delta^{\langle n\rangle}(x_0,s)\dfrac{y_0-1}{h_0-1}, & h_0\leqq y_0\leqq 1,\\ \delta^{\langle n\rangle}(x_0,s)\dfrac{y_0}{h_0}, & 0\leqq y_0\leqq h_0.\end{cases} \tag{7.10}$$

On the free surface $y_0 = h_0$ of the reference domain, we have

$$y^{\langle n\rangle}(x_0, h_0, s) = \delta^{\langle n\rangle}(x_0, s). \tag{7.11}$$

The equations governing the coefficients in (7.6) are very complicated because the differential operators with derivatives with respect to x and y in the field equations must be reexpressed by derivatives with respect to x_0, y_0 under the change of variables $x,\ y \to x_0,\ y_0$. Since this mapping is invertible we can continue (7.6) as

$$\left.\begin{matrix}\mathbf{u}(x,y,s,\varepsilon)\\ \mathbf{v}(x,y,s,\varepsilon)\\ \text{etc.}\end{matrix}\right\} = \sum_{l=0}^{\infty} \frac{\varepsilon^{l+1}}{(l+1)!} \begin{matrix}\mathbf{u}^{[l]}(x, y_0(x,y,s,\varepsilon), s)\\ \mathbf{v}^{[l]}(x, y_0(x,y,s,\varepsilon), s)\\ \text{etc.}\end{matrix} \tag{7.12}$$

Fortunately it is never necessary to solve the complicated equations which govern the derivatives $[l]$ on the right of (7.6). In fact we need only to do much simpler calculations for the partial derivatives $\langle l\rangle$. When the partial derivatives $\langle l\rangle$ are known the total derivatives $[l]$ may be computed by the chain rule (7.9). The point of simplicity of partial derivatives is that they do not perturb the operators which are defined on region Ω_1 and Ω_2, below and above the free surfaces, see [18] and [19]. For example,

$$\operatorname{div} \mathbf{u}^{\langle n\rangle}(x_0, y_0, s) = \frac{\partial u_1^{\langle n\rangle}}{\partial x_0} + \frac{\partial u_2^{\langle n\rangle}}{\partial y_0} = 0, \tag{7.13}$$

whereas

$$\operatorname{div} \mathbf{u}^{[n]}(x_0, y_0, s) \neq 0.$$

The same type of simplification holds for the perturbation equations which arise from (7.1), (5.3) and (5.4). For example,

$$\rho\left[\omega_0 \frac{\partial \mathbf{u}^{\langle 2\rangle}}{\partial s} + \hat{u}_1 \frac{\partial \mathbf{u}^{\langle 2\rangle}}{\partial x} + \mathbf{e}_x u_2^{\langle 2\rangle} \hat{u}_1' + \hat{u}_1^{\langle 2\rangle} \frac{\partial \mathbf{u}^{\langle 0\rangle}}{\partial x} + \mathbf{e}_x u_2^{\langle 0\rangle} \hat{u}_1'^{\langle 2\rangle} + \mathbf{u}^{\langle 1\rangle} \cdot \nabla \mathbf{u}^{\langle 1\rangle} + \omega_2 \frac{\partial \mathbf{u}^{\langle 0\rangle}}{\partial s}\right] = \eta_1 \Delta \mathbf{u}^{\langle 2\rangle} - \nabla p^{\langle 2\rangle}. \tag{7.14}$$

The unknowns here are $\mathbf{u}^{\langle 2\rangle}$, $p^{\langle 2\rangle}$, ω_2 and $\hat{V}_2$.

It is not possible to avoid the total derivatives $[l]$ in (5.6)–(5.10) because these are defined on a manifold, the interface, and not in a region. The interface conditions are of the form

$$g(x, y = h(x,s,\varepsilon), s, \varepsilon) = 0 \tag{7.15}$$

and the perturbation of $y = h$ with ε cannot be avoided. In fact the interface conditions are identities on the interface so that tangential derivatives on them vanish (see [19]).

The following is a recipe for perturbations of the domain in bifurcation problems. First, we introduce the frequency $\omega(\varepsilon)$ into (5.3), (5.4) and (5.10) using (7.1). We then insert the series (7.4), (7.5) and the series

$$\left.\begin{matrix}\mathbf{u}(x_0,y_0,s,\varepsilon)\\ \mathbf{v}(x_0,y_0,s,\varepsilon)\\ p(x_0,y_0,s,\varepsilon)\\ q(x_0,y_0,s,\varepsilon)\\ \delta(x_0,s,\varepsilon)\end{matrix}\right\} = \sum_{l=0} \frac{\varepsilon^{l+1}}{(l+1)!} \begin{matrix}\mathbf{u}^{\langle l\rangle}(x_0,y_0,s),\\ \mathbf{v}^{\langle l\rangle}(x_0,y_0,s),\\ p^{\langle l\rangle}(x_0,y_0,s),\\ q^{\langle l\rangle}(x_0,y_0,s),\\ \delta^{\langle l\rangle}(x_0,s)\end{matrix} \tag{7.16}$$

into $\operatorname{div}\mathbf{u}=\operatorname{div}\mathbf{v}=0$, (5.3), (5.4) and (5.5) and identify the perturbation equations. These equations hold in the reference domain. To get the equations which govern the interface conditions (5.6)–(5.10) we insert the series (7.6) and identify. Then we express the derivatives $[l]$ with partial derivatives $\langle l\rangle$, using the chain rule. The perturbation equations arising from interface conditions are thus defined on the flat interface $y=h_0$.

The series on the right of (7.16) may be expressed on the deformed domain by inverting the mapping, as in (7.12). In fact, the series on the right of (7.16) is equal to the series on the right of (7.6), though the partial sums of these two series are not equal (see equations of [18]).

8. Solvability of the perturbation equations. We must solve perturbation problems of the following form:

(i) All functions of s are 2π periodic in s.

(ii) All functions of $x=x_0$ are L periodic in x_0.

(iii) In the region $0\leqq y_0\leqq h_0$,

$$\frac{\partial u_1^{\langle n\rangle}}{\partial x_0}+\frac{\partial u_2^{\langle n\rangle}}{\partial y_0}=0, \tag{8.1}$$

$$\rho\left[\omega_0\frac{\partial u_1^{\langle n\rangle}}{\partial s}+\hat{u}\frac{\partial \mathbf{u}^{\langle n\rangle}}{\partial x_0}+\mathbf{e}_x u_2^{\langle n\rangle}\hat{u}'\right]-\eta_1\Delta\mathbf{u}^{\langle n\rangle}+\nabla p^{\langle n\rangle}=\theta_1(\omega_n,\hat{V}_n,x_0,s). \tag{8.2}$$

(iv) In the region $h_0\leqq y_0\leqq 1$ we have the same equations with $\mathbf{v}^{\langle n\rangle}(x_0,y_0,s)$, $\hat{v}(y_0)$, $q^{\langle n\rangle}$ and $\boldsymbol{\theta}_2$ replacing $\mathbf{u}^{\langle n\rangle}$, $\hat{u}$, $p^{\langle n\rangle}$ and $\boldsymbol{\theta}_1$.

(v) $u^{\langle n\rangle}=0$ at $y_0=0$, $v^{\langle n\rangle}=0$ at $y_0=1$.

(vi) On the interface $y=h_0+\delta$, we have by (5.8)

$$\delta=k[[u_1]],\ [[u_1]]\overset{\text{def}}{=}u_1-v_1. \tag{8.3}$$

We have eliminated δ from the interface equations (5.6), (5.7) and (5.10) with (8.3). Of course

$$\delta^{\langle n\rangle}=k\left[\left[u_1^{\langle n\rangle}\right]\right]\quad\text{on } y_0=h_0.$$

(vii) The four interface conditions on $y_0=h_0$ are of the form

$$u_2^{\langle n\rangle}-v_2^{\langle n\rangle}=\theta_3(\hat{V}_n,x_0,s),$$

$$\eta_1 D_{12}[\mathbf{u}^{\langle n\rangle}]-\eta_2 D_{12}[\mathbf{v}^{\langle n\rangle}]=\theta_4(\hat{V}_n,x_0,s),$$

$$-p^{\langle n\rangle}+2\eta_1 D_{22}[\mathbf{u}^{\langle n\rangle}]+q^{\langle n\rangle}-2\eta_2 D_{22}[\mathbf{v}^{\langle n\rangle}]-kT\frac{\partial^2}{\partial x_0^2}\left[\left[u_1^{\langle n\rangle}\right]\right]=\theta_5(\hat{V}_n,x_0,s),$$

$$k\left(\omega_0\frac{\partial}{\partial s}+\hat{u}_1(k_0)\frac{\partial}{\partial x_0}\right)\left[\left[u_1^{\langle n\rangle}\right]\right]-u_2^{\langle n\rangle}=\theta_6(\omega_n,\hat{V}_n,x_0,s).$$

The inhomogeneous terms θ are linear in the unknown parameters ω_n and $\hat{V}_n$ and are otherwise known from computations at orders $l<n$.

These inhomogeneous problems can be solved uniquely in the space orthogonal to the null space of the adjoint operator introduced at the begining of §6. This null space is two-dimensional and is spanned by

$$\mathbf{z}^*\quad\text{and}\quad\bar{\mathbf{z}}^*$$

defined by (7.3) and explained in Iooss and Joseph [17, §VIII.3]. There are therefore two solvability conditions to be used in the determination of the parameters ω_n and $\hat{V}_n$.

REFERENCES

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

[2] J. T. BEALE, *Large time regularity of viscous surface waves*, Arch. Rational Mech. Anal., 84 (1984), pp. 307–352.

[3] _____, *The initial value problem for the Navier–Stokes equations with a free surface*, Comm. Pure Appl. Math., 34 (1980), pp. 359–392.

[4] M. G. CRANDALL AND P. H. RABINOWITZ, *The Hopf bifurcation theorem in infinite dimensions*, Arch. Rational Mech. Anal., 67 (1978), pp. 53–72.

[5] C. E. HICKOX, *Instability due to viscosity and density stratification in axisymmetric pipe flow*, Phys. Fluids, 14 (1971), pp. 251–262.

[6] A. P. HOOPER AND W. G. C. BOYD, *Shear-flow instability at the interface between two viscous fluids*, J. Fluid Mech., 128 (1983), pp. 507–528.

[7] E. HOPF, *Abzweigung einer periodischen Lösung eines Differentialsystems*, Ber. Sächs. Akad. Wiss., Math.-nat. Kl., 94 (1942), pp. 1–22.

[8] G. IOOSS, *Existence et stabilité de la solution périodique secondaire intervenant dans les problèmes d'évolution du type Navier–Stokes*, Arch. Rational Mech. Anal., 47 (1972), pp. 301–329.

[9] V. I. IUDOVICH, *The onset of auto-oscillations in a fluid*, J. Appl. Math. Mech., 35 (1971), pp. 587–603.

[10] D. D. JOSEPH AND D. H. SATTINGER, *Bifurcating time periodic solutions and their stability*, Arch. Rational Mech. Anal., 45 (1972), pp. 75–109.

[11] D. D. JOSEPH, M. RENARDY AND Y. RENARDY, *Instability of the flow of two immiscible liquids with different viscosities in a pipe*, J. Fluid Mech., 141 (1984), pp. 309–317.

[12] T. KATO, *Perturbation Theory for Linear Operators*, 2nd ed., Springer, Berlin, 1976.

[13] Y. RENARDY AND D. D. JOSEPH, *Two-fluid Couette flow between concentric cylinders*, J. Fluid Mech., 150 (1985), pp. 381–394.

[14] D. H. SATTINGER, *Bifurcation of time periodic solutions of the Navier–Stokes equations*, Arch. Rational Mech. Anal., 41 (1971), pp. 66–80.

[15] R. TEMAM, *Navier–Stokes Equations*, North-Holland, Amsterdam, 1979.

[16] C. S. YIH, *Instability due to viscosity stratification*, J. Fluid Mech., 27 (1967), pp. 337–352.

[17] G. IOOSS AND D. D. JOSEPH, *Elementary Stability and Bifurcation Theory*, Springer, Berlin, 1980.

[18] D. D. JOSEPH, *Domain perturbations. The higher order theory of infinitesimal water waves*, Arch. Rational Mech. Anal., 51 (1973), pp. 295–303.

[19] _____, *Free surface problems in rheological fluid mechanics*, Rheol. Acta, 16 (1977), pp. 169–189.

SIAM J. MATH. ANAL.
Vol. 17, No. 4, July 1986

011

GLOBAL HOPF BIFURCATION FOR VOLTERRA INTEGRAL EQUATIONS*

BERNOLD FIEDLER†

Abstract. This paper investigates global bifurcation of time periodic orbits for autonomous systems of integral equations of convolution type depending on a real parameter λ. An easy criterion for global bifurcation is derived: if—for simplicity—there exists only one stationary, nondegenerate solution for all λ, then it is sufficient that the linear unstable dimensions for λ near $-\infty$ resp. $+\infty$ differ from each other.

The theorem requires the integral kernels to be integrable with some exponential weight. The proof then relies on approximation by ordinary differential equations.

Applications are given to oscillations in a model for epidemics, and to a model for circular neural nets.

Key words. bifurcation, periodic solutions, integral equations, epidemics, neural nets

AMS(MOS) subject classifications. Primary 45D05, 45L05, 45M99

Introduction. Periodic oscillations are a quite common feature of many models in mathematical biology which take the form of integral equation systems

$$x(t)=\int_0^\infty a_\lambda(s)f_\lambda(x(t-s))\,ds, \quad x\in\mathbb{R}^m, \quad \lambda\in I=[\alpha,\beta] \tag{0.1}$$

see e.g. [10], [11], [16], [22], [33], or of integro-differential systems

$$x'(t)=\int_0^\infty a_\lambda(s)f_\lambda(x(t-s))\,ds, \tag{0.2}$$

see e.g. [4], [8], [9], [22], [23], [26], [30], [31], [34]. A frequent approach to detect such oscillations is local Hopf bifurcation theory with respect to a parameter λ: the stationary (i.e. t-independent) solutions x are computed first and various conditions on the linearized equation for some $\lambda=\lambda_0$ yield small amplitude periodic solutions. Such theorems are available at a considerable level of generality and technical perfection [5], [8], [12], [16], [20], [21], [26], [32]. However, there are two main drawbacks to this approach:

- the assumptions on the linearization are difficult to check, restricting applicability to very special kernels, and
- conclusions are obtained only in a small neighborhood of the bifurcation point.

Taking a global point of view we attack both problems simultaneously, concentrating on integral equation (0.1). We define a global Hopf index H (see Definition 1.6) which requires only some information on the linearizations at $\lambda=\alpha$ and $\lambda=\beta$. Our Main Theorem 3.1 states: if

$$H\neq 0 \tag{0.3}$$

*Received by the editors June 7, 1984, and in revised form February 15, 1985. This work was supported by the Deutsche Forschungsgemeinschaft.

† Universität Heidelberg, Institut für Angewandte Mathematik, D-6900 Heidelberg, Federal Republic of Germany.

then there is a global continuum Z of pairs (λ, x) with x a periodic solution of (0.1). Here "global" means that

- Z is unbounded, or
- the virtual periods of the nonstationary periodic solutions on Z are unbounded, or
- Z contains (λ, x) with x nonstationary periodic, $\lambda = \alpha$ or $\lambda = \beta$.

We call q a virtual period of x, if q is the minimal period of a pair (x, y) where y solves the linearized equation

$$(0.1)' \qquad y(t) = \int_0^\infty a_\lambda(s) f_\lambda'(x(t-s)) y(t-s)\, ds,$$

see Definition 2.1. Note that x may have several virtual periods, but at least the minimal period of x is always a virtual period. Unlike the set of all periods, the set of all virtual periods of x is bounded (Lemma 2.3).

We briefly outline the proof to clarify the role of virtual periods. Our starting point is a corresponding theorem for ordinary differential equations. This theorem relies heavily on ideas of Alligood, Chow, Mallet-Paret and Yorke [2], [3], [7], [27], who also introduced a notion of virtual period. A global theorem involving the Hopf index H was proved by the author [15] in an analytic semigroup setting using generic approximations—a method where virtual periods seem both natural and indispensable. To extend our theory to integral equations we avoid flow concepts like [12], [18] because we were unable to develop a generic theory for equations of the form (0.1). Rather, in §1 we approximate the kernel $a_\lambda(s)$ by exponential sums. For exponential sum kernels (0.1) reduces to an ordinary differential equation and [15] applies. Limits of periodic solutions yield periodic solutions of (0.1). In §2 we establish the crucial fact that limits of virtual periods are again virtual periods. If we replace "virtual" by "minimal", this is no longer true. In §3 we piece everything together and comment on our result.

Of course, there are also some drawbacks to our theorem. In applications, it seems to be difficult to obtain upper bounds on (minimal) periods, in general. Actually this difficulty goes back to the original paper by Alexander and Yorke [0] who started global Hopf bifurcation for ODEs. The problem persists in more recent work [6], [24], [25] and is only slightly diminished in our approach: there is an example of a bifurcation which is global in the sense of [0] but not in ours, see [1]. At present there is one other method to avoid this difficulty, using Poincaré return maps of cones and some fixed point theory, see [4], [19], [20], [30], [31], [34] for example. This method requires detailed information on the nonlinear flow; it works for scalar equations and a very few special systems. Another serious drawback, which we hope to overcome soon, is the following: we cannot allow for stationary bifurcations, because we exclude characteristic value zero of the linearization.

There are many applications of our theorem. For §4, we selected just two to demonstrate the necessary adaptations of our theory and the comparative ease in computing the global Hopf index H. We can also treat integro-differential systems (0.2), but only for kernels which are L^1 with respect to an exponential weight. At present our results do not include delay equations, or the claims of [6].

1. Approximation by ordinary differential equations. In this section we analyze approximations of the integral equation (IE)

$$(1.1) \qquad x(t) = \int_0^\infty a_\lambda(s) f_\lambda(x(t-s))\, ds =: (a_\lambda * f_\lambda(x))(t), \qquad \lambda \in I = [\alpha, \beta]$$

by ordinary differential equations in $\mathbb{R}^n$

$$(1.2)_n \qquad \dot{y} = g_\lambda^n(y),$$

letting the dimension n tend to infinity. In particular we relate eigenvalues of the linearization of the ODE $(1.2)_n$ to the characteristic roots of the linearization of IE (1.1). As a consequence we can define a Hopf index $H(a,f)$ of (1.1) as the limit

$$H(a,f) := \lim_{n\to\infty} H(g^{(n)})$$

of the Hopf indices $H(g^{(n)})$ of equations $(1.2)_n$. Along our way, we develop the technical frame and state a compactness result for later reference.

We fix some notation. For kernels $a_\lambda, b \in L^1(\mathbb{R}^+, M(m,\mathbb{R}))$, a_λ depending continuously on λ, and for $\gamma \geqq 0$, $I=[\alpha,\beta]$ a compact interval, $\alpha<\beta$, we define

$$|b|_\gamma := \int_0^\infty |b(s)| \exp(\gamma s)\, ds, \qquad \|a\|_\gamma := \sup_{\lambda\in I} |a_\lambda|_\gamma,$$

as weighted norms with corresponding spaces

$$L^1_\gamma := \left\{ b \in L^1(\mathbb{R}^+, M(m,\mathbb{R})) \,|\, |b|_\gamma < \infty \right\},$$
$$A_\gamma(I) := \left\{ a : I\times\mathbb{R}^+ \to M(m,\mathbb{R}) \,|\, \|a\|_\gamma < \infty \right\}.$$

We assume that

$$\begin{aligned} &a \in A_\gamma(\mathbb{R}) \text{ with some } \gamma>0 \quad \text{and} \\ &f \in F := C^0\big(I, BC^1(\mathbb{R}^m,\mathbb{R}^m)\big), \quad \text{i.e.,} \\ &f_\lambda \in BC^1(\mathbb{R}^m,\mathbb{R}^m) \text{ depends continuously on } \lambda\in I. \end{aligned} \tag{1.3}$$

Here C^k (BC^k) denotes the space of functions with (uniformly bounded) continuous derivatives up to order k, endowed with the standard sup-norm. By C_0, we will denote continuous functions with compact support. As a solution space of equation (1.1) we consider

$$x \in X := BC^0(\mathbb{R},\mathbb{R}^m).$$

We call $x\in X$ periodic if $x(t+p)=x(t)$ for some $p>0$ and all real t, including stationary solutions which admit any real p as a period. By assumption (1.3), all periodic solutions of (1.1) are in X. It is not necessary to introduce a phase space of past histories for IE (1.1) if we concentrate on periodic solutions only. Thus, we avoid the technicalities of defining semiflows for (1.1) but we also lose concepts like Poincaré-maps and Floquet-multipliers.

Let us state compactness. Using [28] we know the following.

LEMMA 1.1. *Let $y_n \in X$ of period p_n satisfy the (not necessarily autonomous) equation*

$$(1.4)_n \qquad y_n(t) = b^n * g^n(\cdot, y_n)(t) := \int_0^\infty b^n(s)\, g^n(t-s, y_n(t-s))\, ds$$

and assume

(i) *b^n converges to b in L^1;*

(ii) *g^n converges to g in $BC^0(\mathbb{R}\times\mathbb{R}^m,\mathbb{R}^m)$ and $g^n(\cdot,y)$ are periodic with periods independent of y and uniformly bounded with respect to n;*

(iii) *p_n converges to $p_\infty<\infty$.*

Then (y_n) has a subsequence converging in X to y with period p_∞ and y satisfies the limiting equation

$$(1.4)_\infty \qquad y = b * g(\cdot, y). \qquad \square$$

We turn to approximation of kernels by exponential sums. Let e_j be the exponential $e_j(s) := \exp(-js)$, and $P_\gamma(I) \subset A_\gamma(I)$ be the subset of finite sums of functions

$$(\lambda, s) \to p_j(\lambda) e_{\gamma+j}(s)$$

with $p_j \in C^0(I, M(m, \mathbb{R}))$, $j \in \mathbb{N}$.

LEMMA 1.2. *$P_\gamma(I)$ is dense in $A_\gamma(I)$.*

Proof. Because multiplication by e_γ is an isomorphism from L^1 to L^1_γ, we prove only the case $\gamma = 0$. Further we may restrict our attention to $m = 1$. I is compact and we allow for continuous $p_j(\lambda)$, hence, fixing λ, we only have to prove: for any $b \in L^1(\mathbb{R}^+, \mathbb{R})$ and for any $\varepsilon > 0$ there exist real coefficients p_j, $j = 1, \cdots, n$ such that $p := \sum p_j e_j$ satisfies

$$|b - p|_{L^1(\mathbb{R}^+, \mathbb{R})} < \varepsilon. \tag{1.5}$$

To prove (1.5), we note that $C_0(\mathbb{R}^+, \mathbb{R})$ is dense in $L^1(\mathbb{R}^+, \mathbb{R})$; hence we may assume that b is continuous and has compact support in $\mathbb{R}^+$. Let

$$b^*(\sigma) := \sigma^{-1} b(-\log \sigma) \in C^0([0,1], \mathbb{R})$$

with the obvious definition $b^*(0) := 0$. By the Stone-Weierstraß approximation theorem, there exists a polynomial $p^*(\sigma) := \sigma^{-1} \sum p_j \sigma^j$ such that

$$\max_{\sigma \in [0,1]} |b^*(\sigma) - p^*(\sigma)| < \varepsilon.$$

For $s = -\log \sigma$ this implies

$$|b(s) - p(s)| = |\sigma b^*(\sigma) - \sigma p^*(\sigma)| < \varepsilon \exp(-s)$$

and integration over $s \in \mathbb{R}^+$ yields (1.5). □

In the next lemma we show that IE (1.1) is equivalent to an ODE, if the kernel a is an exponential sum (i.e. if $a \in P_\gamma(I)$).

LEMMA 1.3. *Let* $a_\lambda = \sum p_j(\lambda) e_{j+\gamma} \in P_\gamma(I)$, $j = 1, \cdots, n$, $\xi = (\xi_1, \cdots, \xi_n)$, $g^n_\lambda = (g_{1,\lambda}, \cdots, g_{n,\lambda})$ *with*

$$g_{j,\lambda}(\xi) := p_j(\lambda) f_\lambda(\xi_1 + \cdots + \xi_n) - (j+\gamma)\xi_j. \tag{1.6}$$

Define the respective transformations

$$x := \sum \xi_j, \qquad \xi_j := p_j(\lambda) \cdot (e_{j+\gamma} * f_\lambda(x)). \tag{1.7}$$

Then, for each $\lambda \in I$, $x \in X$ is a solution of IE (1.1) *iff ξ is a bounded solution of* ODE $(1.2)_n$.

The proof consists of an obvious calculation and is omitted. □

Combining Lemmata 1.1–1.3 the following proposition is immediate.

PROPOSITION 1.4. *Let a and f satisfy assumption* (1.3). *Then, for each $\lambda \in I$, there is a sequence $a^n_\lambda \in P_\gamma(I)$ of exponential sum kernels, converging to a_λ in $A_\gamma(I)$, such that solutions $x^n \in X$ of*

$$x^n = a^n_\lambda * f_\lambda(x^n) \tag{$1.1)_n$}$$

correspond by transformation (1.7) *to bounded solutions ξ^n of*

$$\dot{\xi}^n = g^n_\lambda(\xi^n), \tag{$1.2)_n$}$$

and the stationary (t-independent) solutions of $(1.1)_n$ *and* (1.1) *are the same.*

In particular, if ξ^n *are periodic solutions of* $(1.2)_n$ *with periods* p_n *converging to* $p_\infty \in \mathbb{R}^+$, *then there is a subsequence of the* x^n *converging in* X *to a periodic solution* x *of* (1.1) *with period* p_∞. □

Now we investigate the behavior of eigenvalues resp. characteristic roots under the approximation from Proposition 1.4. Let $x(t) \equiv x_0$ be a stationary solution of (1.1), then the linearized equation is

$$(1.1)' \qquad y = a_\lambda * \big(f'_\lambda(x_0)\cdot y\big).$$

Values $\mu \in \mathbb{C}$, such that (1.1)′ has a nontrivial solution

$$y(t) = e^{\mu t} y_0,$$

are called *characteristic roots* of (λ, x_0). Let $\hat{a}_\lambda(\mu) := \int_0^\infty \exp(-\mu s) a_\lambda(s)\, ds$ denote the Laplace transform of a_λ (defined for $\operatorname{Re}\mu \geqq -\gamma$); then characteristic roots (with $\operatorname{Re}\mu > -\gamma$) are the zeros of the characteristic function

$$(1.8) \qquad \chi(\mu) = \det\big(1 - \hat{a}_\lambda(\mu)\cdot f'_\lambda(x_0)\big).$$

By analyticity of $\chi(\mu)$, we may assign a multiplicity to each characteristic root μ. For $a_\lambda \in P_\lambda(I)$, linearization commutes with the equivalence transformation from IE (1.1) to ODE (1.2), by Lemma 1.3. Hence we may also associate the linearization

$$(1.2)' \qquad \dot{\eta} = g'_\lambda(\xi_0)\eta$$

to (λ, x_0) and consider the eigenvalues of $g'_\lambda(\xi_0)$, where ξ_0 corresponds to x_0 by (1.7).

LEMMA 1.5. *In the situation described above, the characteristic roots* μ *of* (1.1) *at* (λ, x_0) *with* $\operatorname{Re}\mu > -\gamma$ *are exactly the eigenvalues* μ' *of* (1.2)′ *with* $\operatorname{Re}\mu' > -\gamma$, *with equal algebraic multiplicities.*

Proof. The considerations above prove the lemma except for the claim that algebraic multiplicities are equal. To complete the proof, we directly compute χ and the characteristic polynomial π of $g'_\lambda(\xi_0)$. Absorbing $f'_\lambda(x_0)$ into a_λ, we may assume $f'_\lambda(x_0) = id$. Further, we may omit the index λ.

With $a = \sum p_j e_{j+\gamma}$, $q_j(\mu) := (j+\gamma+\mu)^{-1} p_j = p_j \cdot \hat{e}_{j+\gamma}$ we first obtain

$$\chi(\mu) = \det\big(1 - \textstyle\sum q_j(\mu)\big).$$

Next we compute π, doing some determinant manipulations:

$$\begin{aligned}
\pi(\mu) &= \det\big(g'(\xi_0) - \mu\big) = \det\big(p_i - (j+\gamma+\mu)\delta_{ij}\big)_{i,j} \\
&= \prod_k (k+\gamma+\mu)\cdot \det\big(q_i - 1\cdot\delta_{ij}\big)_{i,j} \\
&= \pm \prod_k (k+\gamma+\mu)\cdot \det\Big(\sum_j q_j - 1\Big) \\
&= \pm \prod_k (k+\gamma+\mu)\cdot \chi(\mu),
\end{aligned}$$

and the lemma is proved. □

Finally, we introduce the Hopf index, using our ODE approximations. First we need some more notation and assumptions. Let (λ_0, x_0) be a stationary solution of IE (1.1) with corresponding linearization (1.1)′ and characteristic function χ; cf. (1.8). We assume

$$(1.9) \qquad \chi(0) \neq 0 \text{ at any stationary solution } (\lambda_0, x_0).$$

In particular, this excludes secondary bifurcation of stationary solutions; the stationary solutions are given as branches $(\lambda, x^j(\lambda))$, parametrized by λ with j labeling the branches. On occasion, we also denote the characteristic function at $(\lambda, x^j(\lambda))$ by $\chi^j_\lambda(\mu)$. Further we call (λ_0, x_0) a *center* if $\chi = 0$ has pure imaginary roots; let $C \subset I \times X$ denote the set of centers of (1.1). We assume

(1.10) the set C of centers does not contain points (λ, x) with $\lambda \in \partial I = \{\alpha, \beta\}$.

Let

$E^j(\lambda)$: number of characteristic roots with positive real part at $(\lambda, x^j(\lambda))$, counting multiplicities.

DEFINITION 1.6. We define the global Hopf index of integral equation (1.1) to be

$$H = H(I) := \frac{1}{2} \cdot \sum_j (-1)^{E^j(\alpha)} \cdot (E^j(\beta) - E^j(\alpha)). \tag{1.11}$$

Choose an exponential approximation of a by $a^n \in {}_\gamma(\mathbb{R})$ such that

$$a^n \underset{n}{\to} a \quad \text{in } A_\gamma(I)$$

and Proposition 1.4 holds. We define $H_n(I)$ by (1.11), corresponding to a^n, for $n \geqq n_0$ large enough that assumptions (1.9) and (1.10) above hold for equation $(1.1)_n$, too. By Lemma 1.5 we may equivalently interpret the characteristic roots contributing to $E^j_n(\lambda)$ for a^n as eigenvalues of the associated ODE.

PROPOSITION 1.7. *Under assumptions* (1.3), (1.9) *and* (1.10), *the global Hopf index is well-defined and satisfies*

$$H(I) = \lim_{n \to \infty} H_n(I).$$

Proof. By definition, we only have to show

$$E(\lambda) = \lim_{n \to \infty} E_n(\lambda),$$

at any stationary solution (λ, x_0) of (1.1), $\lambda \in \partial I$ (by Proposition 1.4, (1.1) has the same stationary solutions as $(1.1)_n$). Suppressing the argument λ, we note that E_n is the number of roots of

$$\chi_n = \det(1 - \hat{a}_n(\cdot))$$

with positive real part, if we put $f'_\lambda(x_0) = 1$ without loss of generality. For any $\varepsilon > 0$, by continuity, there exists $n_0(\varepsilon)$ such that for all $n \geqq n_0(\varepsilon)$

$$|\chi_n(\mu) - \chi(\mu)| < \varepsilon \quad \text{for } \operatorname{Re}\mu \geqq 0.$$

Now $\chi \neq 0$ on the imaginary axis, and $\lim_{|\mu| \to \infty} \chi(\mu) = 1$ for $\operatorname{Re}\mu \geqq 0$, by the Riemann–Lebesgue lemma, hence zeros of χ in $\{\operatorname{Re}\mu \geqq 0\}$ are in some compact subset of $\{\operatorname{Re}\mu > 0\}$. By uniform convergence of the analytic functions χ_n to χ,

$$E_n = E$$

for $\varepsilon > 0$ small enough and $n \geqq n_0(\varepsilon)$, and the proof is complete. □

2. Virtual periods. We develop a notion of virtual periods for periodic solutions of integral equations, without referring to flow concepts, Poincaré maps etc.. Our definition is consistent with the virtual periods introduced by Alligood, Mallet-Paret and

Yorke [2], [3], [27] in an ODE setting. We show that virtual periods are well behaved under limits, i.e. the limit of virtual periods is again a virtual period. Then we prove that any orbit has only finitely many virtual periods, and finally we analyze the behavior of virtual periods on branches of stationary solutions, using analyticity. Throughout this section let assumptions (1.3) on a, f and (1.9): $\chi(0)\neq 0$ be satisfied.

DEFINITION 2.1. Given a periodic solution $x(\cdot)$ of integral equation

$$x = a * f(x) \tag{1.1}$$

we call $q>0$ *a virtual period of* x if q is the prime (minimal) period $p(x,y)$ of a pair (x,y), where y is a periodic solution of the linearized equation

$$y = a * (f'(x(\cdot))y). \tag{1.1}'$$

Recall that we regard stationary solutions as periodic with "minimal period" $p := 0$. Setting $y := 0$, the prime period is always a virtual period; however, $x(\cdot)$ may have several virtual periods (see Proposition 2.2 below). Floquet-theory, along with Lemma 1.3 on the correspondence between ODEs and IEs, tells us that our definition of virtual periods coincides with the one given in [2], [3], [27] for ODEs, if $a\in P_\gamma$ is an exponential sum kernel.

In [7], [27] it was shown that a limit of *minimal* periods is a *virtual* period. The most useful property of virtual periods in our approach is a generalization of this result: we prove that a limit of *virtual* periods is again a *virtual* period. Because of the independent significance of the next proposition, we emphasize that the proof does not use Assumption (1.9) $\chi(0)\neq 0$; but this fact is not exploited in the present paper.

PROPOSITION 2.2. *Let* x_n *be periodic solutions of*

$$x_n = a_n * f_n(x_n) \tag{1.1}_n$$

with a virtual period $q_n>0$. *For* $n\to\infty$ *assume that*

$$\begin{aligned} a_n &\to a && \text{in } L^1,\\ f_n &\to f && \text{in } BC^1,\\ x_n &\to x && \text{in } X = BC^0,\\ q_n &\to q_\infty && \text{in } \mathbb{R}. \end{aligned}$$

Then q_∞ *is a virtual period of* x, *in particular* q_∞ *is positive.*

Proof. The proof is rather involved; we give an outline first. We construct two solutions η_1 and η_2 of the linearized equation (1.1)′ such that (x,η_1,η_2) has minimal period $p(x,\eta_1,\eta_2)=q_\infty>0$. This is sufficient because $p(\eta_1,\eta_2)=p(\alpha_1\eta_1+\alpha_2\eta_2)$ for almost every $\alpha_1,\alpha_2\in\mathbb{R}$. In the first step we construct η_1, involving limits of x_n, such that

$$p(x,\eta_1)=p_\infty,$$

where $p_\infty=\lim p_n$, $p_n := p(x_n)$. In our second step we construct η_2. This involves limits of solutions y_n of the linearized equation (1.1)′ with $p(x,y_n)=q_n$, $|y_n|=0$ or 1. Finally we prove that indeed

$$p(x,\eta_1,\eta_2)=q_\infty>0.$$

Note that in our proof we may assume that $p_n>0$ eventually; otherwise, for x_n stationary, we consider the linearized equations for y_n directly replacing x_n by y_n in the proof.

For later reference we introduce some notation. For periodic $\phi \in X$ we denote

$p(\phi)$: the minimal period of ϕ,
$P(\phi)$: the set of all periods of ϕ,
$S(t)\phi := \phi(\cdot + t)$, the shift on X,
$\delta_n^\sigma := |(S(\sigma p_n)x_n - x_n)|_X > 0$ for $p_n > 0$, $\sigma \in \mathbb{Q} \setminus \mathbb{Z}$,
$x_n^\sigma := (S(\sigma p_n)x_n - x_n)/\delta_n^\sigma$,
$z_n := (x_n, y_n)$,
$\varepsilon_n^\tau := |S(\tau q_n)z_n - z_n|_{X \times X} > 0$ for $\tau \in \mathbb{Q} \setminus \mathbb{Z}$,
$y_n^\tau := (S(\tau q_n)y_n - y_n)/\varepsilon_n^\tau$.

Step 1: *Construction of* η_1. We claim that, after restriction to an appropriate subsequence,

$$x^\sigma := \lim_n x_n^\sigma \tag{2.1}$$

exists and satisfies the linearized equation (1.1)′ at x, provided that $\sigma \in \mathbb{Q} \setminus \mathbb{Z}$ satisfies the admissibility condition

$$\sigma p_\infty \in P(x). \tag{2.2}$$

For admissible σ (possibly an empty set) we then show

$$\sigma p_\infty \notin P(x^\sigma). \tag{2.3}$$

We choose η_1 as a finite linear combination

$$\eta_1 := \sum_\sigma \alpha_\sigma x^\sigma$$

with admissible σ, such that $p(\eta_1) = p(x^\sigma, \sigma \text{ admissible})$ and we finally show

$$p(x, \eta_1) = p_\infty.$$

Existence of x^σ follows from our compactness Lemma 1.1: by admissibility (2.2) we have $\lim_n \delta_n^\sigma = 0$ and therefore

$$x_n^\sigma = a_n * (f_n'(x_n)x_n^\sigma) + o(1),$$

which yields convergence and

$$x^\sigma = a * (f'(x)x^\sigma).$$

We prove indirectly that $\sigma p_\infty \notin P(x^\sigma)$: choose $k \in \mathbb{N}$ such that $k\sigma \in \mathbb{Z}$ and compute

$$\begin{aligned} 0 &= \lim_n (S(k\sigma p_n)x_n - x_n)/\delta_n^\sigma \\ &= \lim_n \sum_{j=0}^{k-1} (S((j+1)\sigma p_n)x_n - S(j\sigma p_n)x_n)/\delta_n^\sigma \\ &= \sum S(j\sigma p_\infty)x^\sigma = kx^\sigma, \end{aligned}$$

if we assume $\sigma p_\infty \in P(x^\sigma)$ in the last equality. This is a contradiction to $|x^\sigma|_X = 1$. Note that (2.3) implies $p_\infty > 0$, hence $q_\infty > 0$. Further, x^σ cannot be stationary.

We show that $p(x,\eta_1)=p_\infty$. If there is no admissible σ, then $\sigma p_\infty \in P(x)$ implies $\sigma \in \mathbb{Z}$, i.e. p_∞ is the minimal period of x and we may choose $\eta_1 \equiv 0$. If there are admissible σ then (2.2), (2.3) imply that

$$\mathbb{Z}p_\infty = P(x) \cap \bigcap_{\sigma \text{ adm}} P(x^\sigma). \tag{2.4}$$

To see that, note that p_∞ is contained in the right-hand side by construction. Vice versa: $\mathbb{Q}p_\infty$ contains the right-hand side; but if $\alpha p_\infty \in P(x) \cap \bigcap P(x^\sigma)$, $\alpha \in \mathbb{Q}$, then $\alpha \in \mathbb{Z}$, or else $\alpha \in \mathbb{Q}\setminus\mathbb{Z}$ is admissible and therefore $\alpha p_\infty \notin P(x^\alpha)$: a contradiction. This proves (2.4). Now $P(x^\sigma)$ is discrete for all σ, therefore a finite intersection over σ is sufficient in (2.4). The corresponding σ provide η_1, as desired.

Step 2: *Construction of* η_2. This step is somewhat similar to the last one, replacing x_n by y_n. We claim that

$$y := \lim y_n, \qquad y^\tau := \lim y_n^\tau \tag{2.1}'$$

exist, y satisfies (1.1)′, and y^τ satisfies (1.1)′ if $\tau \in \mathbb{Q}\setminus\mathbb{Z}$ is admissible, i.e.

$$\tau q_\infty \in \mathbb{Z}p_\infty \cap P(z) \tag{2.2}'$$

where $z=(x,y)$. For admissible τ we show

$$\tau q_\infty \notin P(y^\tau) \tag{2.3}'$$

and choose η_2 as a finite linear combination of z^τ such that finally

$$p(x,\eta_1,\eta_2) = q_\infty.$$

Convergence of y_n is trivial. However, some care is needed with y^τ to insure that y^τ satisfies the linearization (1.1)′; this explains the $\mathbb{Z}p_\infty$ in condition (2.2)′ which in turn made step 1 indispensable. By $\tau q_\infty \in P(z)$ the ε_n^τ go to zero. In addition $\tau q_\infty \in \mathbb{Z}p_\infty$, $p_\infty > 0$ implies that $\tau q_n \in \mathbb{Z}p_n$, eventually. Therefore,

$$S(\tau q_n)x_n = x_n, \quad \varepsilon_n^\tau = |S(\tau q_n)y_n - y_n|_X, \quad |y_n^\tau| = 1$$

and we may compute that

$$\begin{aligned} y_n^\tau &= a_n * \big(S(\tau q_n)\big(f_n'(x_n)y_n\big) - f_n'(x_n)y_n\big)/\varepsilon_n^\tau \\ &= a_n * \big(f_n'(S(\tau q_n)x_n)S(\tau q_n)y_n - f_n'(x_n)y_n\big)/\varepsilon_n^\tau \\ &= a_n * \big(f_n'(x_n)y_n^\tau\big) \end{aligned}$$

indeed satisfies the linearization (1.1)′.

We prove (2.3)′ as in Step 1, replacing (p,x,δ,σ) by (q,y,ε,τ).

Finally we show that $p(x,\eta_1,\eta_2)=q_\infty$. If there is no admissible τ, then

$$\tau q_\infty \in \mathbb{Z}p_\infty \cap P(z)$$

implies $\tau \in \mathbb{Z}$ and we may choose $\eta_2 = y$ to obtain $p(x,\eta_1,y)=q_\infty$. If there are admissible τ, (2.2)′ and (2.3)′ imply

$$\mathbb{Z}q_\infty = \mathbb{Z}p_\infty \cap P(z) \cap \bigcap_{\tau \text{ adm}} P(y^\tau)$$

as before. By $p_\infty > 0$, there are only finitely many admissible $\tau \in (0,1)$ anyway. Choosing η_2 as an appropriate linear combination of the corresponding y^τ and y completes the proof. □

In view of Lemma 1.3 on the correspondence between integral equations and ordinary differential equations, the virtual period Proposition 2.2 also holds for ODEs, of course.

In §3 we consider global bifurcation and one alternative will be that arbitrarily large virtual periods occur. Such a statement is significant only if we show that for any single orbit the virtual periods are bounded.

LEMMA 2.3. *Let x be a periodic or stationary solution of*

$$x = a * f(x). \tag{1.1}$$

Then the virtual periods of x are a bounded set.

Moreover, suppose x is stationary. Then x has a (positive) virtual period iff x is a center. More precisely: q is a virtual period of x, iff q is the (positive) least common multiple of numbers $2\pi/\beta_k$, where $\{\pm i\beta_\nu \mid \nu = 1, \cdots, j\}$ are the pure imaginary characteristic roots of x and k ranges over a subset of $\{1, \cdots, j\}$. In particular, x has at most $2^j - 1$ distinct virtual periods.

Proof. Let y be a periodic solution of the linearized equation

$$y = a * f'(x)y \tag{1.1}'$$

such that p is the period of x and $q = p(x,y)$ is a virtual period of x. The case $p = 0$, i.e. x stationary, is elementary and well be treated first. For nonstationary x, we use that (1.1)′ defines a linear evolution system on an exponentially weighted history space, i.e. we rely on a (linear) flow concept. The evolution will be α-condensing, thus proving boundedness of the virtual periods.

For stationary x, let

$$y(t) = \sum_{r \in \mathbb{Z}} y_r \exp(ir\beta t) \qquad \beta := 2\pi/q,$$

be the Fourier series for y. From (1.1)′ we see that

$$y_r = 0 \quad \text{or} \quad \chi(ir\beta) = 0.$$

Hence r shows up in the Fourier series, iff $r\beta = \pm\beta_k$ for some k and the claims are immediate.

For x with period $p > 0$, let Y be the set of continuous functions

$$y_0 : (-\infty, 0] \to \mathbb{R}^m,$$

such that for γ as in (1.3)

$$\|y_0\|_\gamma := \sup_{s \leqq 0} |y_0(s)| \exp(\gamma s) < \infty,$$

and let Y_s be the set of $y_0 \in Y$ such that

$$y_0(0) = a * \big(f'(S(s)x)\, y_0\big)(0).$$

Obviously, Y and Y_s are Banach spaces and (1.1)′ defines a solution operator

$$U(t,s) : Y_s \to Y_t,$$

see [28]. Because x is p-periodic, $T := U((j+1)p, jp)$ is independent of j, maps Y_0 into itself, and

$$y = S(q)y = T^n y, \tag{2.5}$$

if $n := q/p$ and y is interpreted, by restriction, as an element of Y_0.

We claim that the spectral radius r_e of the essential spectrum of T is at most $\exp(-p\gamma) < 1$. Then y lies in the generalized eigenspaces of T for some eigenvalues which are nth roots of unity. But T has at most finitely many eigenvalues on the unit circle; hence n and consequently q is bounded.

To prove the claim, we apply Nussbaum's formula [29] which states

$$r_e = \lim_{n\to\infty} (\alpha(T^n))^{1/n},$$

where $\alpha(T^n)$ is the Hausdorff measure of noncompactness of $T^n(B_1)$. To compute α we split T^n into a sum

$$T^n = S_n + K_n,$$

where S_n is roughly a shift and K_n is compact. To be specific, we define

$$(S_n y_0)(t) := \begin{cases} y_0(t+np), & -\infty < t \leqq -np, \\ (-n+1-t/p)\, y_0(0) & \text{for } -np \leqq t \leqq (-n+1)p, \\ 0, & (-n+1)p \leqq t \leqq 0, \end{cases} \tag{2.6}$$

$$(K_n y_0)(t) := \begin{cases} 0, & -\infty < t \leqq -np, \\ w(t+np) - (-n+1-t/p)\, y_0(0) & \text{for } -np \leqq t \leqq (-n+1)p, \\ w(t+np), & (-n+1)p \leqq t \leqq 0, \end{cases}$$

where w is the solution of the Volterra equation

$$w(t) = \int_0^t a(t-s) f'(x(s)) w(s)\, ds + h(t),$$

$$h(t) := \int_{-\infty}^0 a(t-s) f'(x(s))\, y_0(s)\, ds.$$

Note that w exists in $BC(\mathbb{R}_0^+)$ and defines the continuation of $y_0 \in Y_0$ as a solution of (1.1)', hence $S_n + K_n = T^n$. Unfortunately, S_n and K_n map into Y, not necessarily into Y_0. But choosing a projection P_0 onto Y_0 (which has codimension 1 in Y), we may write

$$T^n = P_0 S_n + P_0 K_n.$$

By compactness of K_n [28], and because $\operatorname{codim} Y_0 = 1$, we may compute

$$\alpha(T^n) = \alpha(P_0 S_n) = \alpha(S_n)$$

[13], [29]. For S_n we estimate

$$\begin{aligned} |S_n y_0|_\gamma &= \sup_{t\leqq 0} |(S_n y_0)(t)| \cdot \exp(\gamma t) \\ &\leqq \max_{\tau\in[0,p]} (1-\tau/p) \exp(\gamma\tau) \cdot \exp(-\gamma np) \cdot |y_0|_\gamma \\ &= c \exp(-\gamma pn) \cdot |y_0|_\gamma, \end{aligned}$$

this implies $\alpha(T^n) \leqq c \exp(-\gamma p n)$, hence

$$r_e = \lim_n \alpha(T^n)^{1/n} \leqq \exp(-p\gamma) < 1,$$

and the proof is complete. □

We conclude this section with a technical lemma on virtual periods of branches of stationary solutions. Recall that $\chi_\lambda^j(\mu)$ denotes the characteristic function (1.8) at the jth stationary branch $(\lambda, x^j(\lambda))$. We assume that

(2.7) for each j, the map

$$(\lambda, \mu) \to \chi_\lambda^j(\mu)$$

is analytic for $(\lambda, x^j(\lambda))$ in a neighborhood of the set C of centers and for $\operatorname{Re}\mu > -\gamma$.

LEMMA 2.4. *Let Assumption* (2.7) *be satisfied. Then there exists a constant* c_0 *such that the set*

$$J(c_0) := \{(\lambda, x) \in C \mid \textit{all virtual periods of } (\lambda, x) \textit{ are } < c_0\}$$

is dense in C.

Proof. For simplicity of notation we focus on just one stationary branch $(\lambda, 0)$ and think of C as a subset of $\mathbb{R}$. In [15, Lemma 4.8] we gave a proof which holds for ODEs, where

$$\chi_\lambda(\mu) = \det(\mu - f_\lambda'(0)).$$

We use Lemma 2.3 to rephrase $J(c_0)$ in terms of zeros of χ. By the Riemann–Lebesgue lemma, the imaginary parts of characteristic values are bounded on C. The only difference to the ODE case is the fact that χ_λ is not polynomial in μ any more. We remedy this drawback locally by application of the Weierstraß preparation theorem, leaving the zeros unchanged. Then we use the proof of [15, Lemma 4.8] to complete our proof. □

3. Global Hopf bifurcation. In this section we present our main abstract result, i.e. we obtain global bifurcation of periodic orbits if the Hopf index H of Definition 1.6 is nonzero

(3.1) $$H \neq 0.$$

We discuss our result and indicate some easy generalizations.

MAIN THEOREM 3.1. *Let* a *and* f *satisfy the following assumptions*:

(1.3) $a \in A_\gamma$ *for some positive exponential weight* γ *and* $f \in F$ (*differentiability*);

(1.9) *the stationary branches are parametrized over* λ *in the compact interval* $I = [\alpha, \beta]$;

(1.10) *there are no centers* (λ, x) *with* $\lambda = \alpha$ *or* $\lambda = \beta$;

(2.7) *the characteristic function* $\chi_\lambda(\mu)$ *is analytic in* (λ, μ);

and—the most important assumption—

(3.1) *the Hopf index* H *is nonzero.*

Then there exists a continuum $Z \subset I \times X$ *consisting of stationary centers and nonstationary periodic solutions* (λ, x) *of*

(1.1) $$x = a_\lambda * f_\lambda(x)$$

such that Z contains both a center and nonstationary periodic solutions and Z is global, i.e. Z contains nonstationary periodic solutions (λ, x) with

(3.2) *arbitrarily large norm of $x \in X$, or*

(3.3) *arbitrarily large virtual period of $x \in X$, or*

(3.4) $\lambda = \alpha$, *or* $\lambda = \beta$ *on the boundary of* I.

We postpone the proof to give a discussion first. By Definition 2.1 virtual periods are some multiples of the minimal period; however, for each (λ, x) in Z the virtual periods of (λ, x) are a finite bounded set (Lemma 2.3). It seems unlikely that we may replace "virtual period" by "minimal period" in conclusion (3.3), except for ODEs of dimension $\leqq 4$. Note that we do not require the periods to become unbounded in a continuous fashion, in contrast to [0], [6], [24], [25]. Actually we have to allow for jumps of the virtual periods by factors of 2 as becomes clear from the proof for ODEs (see e.g. [15, Thm. 4.7]). Technically, the advantage of virtual periods is the fact that limits of virtual periods are again virtual periods (Proposition 2.2).

Using Proposition 2.2 for stationary limiting x only, we could also obtain a global Hopf bifurcation result in the spirit of Alexander and Yorke [0] by our approximation process. In contrast to their result, we do not require any detailed analysis at the centers (e.g. concerning multiplicities of pure imaginary eigenvalues, resonance conditions etc.). All the necessary information sits in the index H. In [0], a bounded continuum of periodic orbits is also called "global", if it just connects centers. Our conclusion above is stronger: we do not call such a continuum global.

By Assumption (1.9) we exclude bifurcations within the class of stationary solutions. This severe drawback already shows up in the ODE case. It is connected to the fact that the index $H = H(f)$ is not homotopy invariant with respect to the families f. We will analyze this problem in a future paper, concentrating on the ODE situation.

Analyticity of the characteristic function χ may be dropped and the proof simplifies, if we assume e.g. that the set C of centers is discrete. In our abstract result, however, we want to use only minimal information on C.

For technical convenience we require f to be bounded and I to be compact. We can drop both restrictions, if we replace Assumption (1.10) by

(1.10)′ the set C of centers of f is bounded in $\mathbb{R} \times X$.

We sketch the necessary modifications. Let $\psi_n \in C^1(\mathbb{R}, \mathbb{R})$ be strictly increasing, $\psi_n(\rho) = \rho$ for $|\rho| \leqq n-1$ and $\psi_n(\rho) = n$ for $|\rho| \geqq n$. Define cut-off approximations f_n of $f: \mathbb{R} \times \mathbb{R}^m \to \mathbb{R}^m$ by

$$f_n(\lambda, x) := f\left(\psi_n(\lambda),\ \psi_n(|x|) \cdot \frac{x}{|x|}\right).$$

By Assumption (1.10)′ the centers C_n of f_n coincide with C for n large enough. Applying Theorem 3.1 to f_n, $I_n := [-n, n]$, we obtain global periodic continua $Z_n \subset I_n \times X$ bifurcating from $C_n = C$. In the proof of our theorem (Step 2) we will indicate how to obtain a limiting global continuum $Z \subset \mathbb{R} \times X$ of f from Z_n as $n \to \infty$. Here global means that Z contains both a center and nonstationary periodic solutions (λ, x) with arbitrarily large

(3.2)′ norm in $\mathbb{R} \times X$ or

(3.3)′ virtual period of $x \in X$.

The variants $I = [\alpha, \infty)$, $I = (-\infty, \beta]$ can be treated similarly.

Our theorem holds for various other autonomous integral equations of convolution type as well. Some examples are

$$x = g(a_1 * f_1(x), \cdots, a_k * f_k(x)),$$
$$\dot{x} = a * f(x),$$

and various other combinations. We just have to give a technical frame, where compactness (Lemma 1.1) and an ODE approximation (Proposition 1.4) holds. However, it is not clear how our results extend to abstract integral equations, e.g. in Hilbert space.

Proof (*Main Theorem* 3.1). We give an outline first. We approximate the original equation

$$x = a_\lambda * f_\lambda(x) \tag{1.1}$$

by integral equations with exponential sum kernels

$$x = a_\lambda^n * f_\lambda(x) \tag{1.1}_n$$

which in turn are equivalent to ODEs

$$\dot{y} = g_\lambda^n(y), \tag{1.2}_n$$

as summarized in Proposition 1.4. In Step 1 we use a global bifurcation result for ODEs (cf. [15]), to construct large continua Z_k^n of periodic solutions of $(1.1)_n$; these continua are getting more and more "global" as $k \in \mathbb{N}$ increases. In Step 2 we pass to the limit $n \to \infty$ and obtain continua Z_k of periodic solutions of the original equation (1.1). Step 3 shows that Z_k is not contained in the set of centers C for k large enough. Defining

$$Z := \bigcup_k Z_k$$

will finish the proof.

Step 1: *properties of* Z_k^n. We consider $k \geqq k_0$ with k_0 large enough that B_{k_0} contains C. Then, by Proposition 1.7, the Hopf index $H_n(k)$ of ODE $(1.2)_n$ on B_k satisfies

$$0 \neq H = \lim_{n \to \infty} H_\nu(k) = H_n(k) \quad \text{for } n \geqq n_0(k),$$

where, for any k, $B_k := \{(\lambda, x) \in I \times X \mid |x|_X \leqq k\}$ with boundary $\partial B_k := \{(\lambda, x) \in I \times X \mid \lambda = \alpha \text{ or } \lambda = \beta \text{ or } |x| = k\}$. Consider $k \geqq k_0$ with k_0 large enough, such that B_{k_0} contains all stationary solutions of f and C is contained in the interior $B_{k_0} - \partial B_{k_0}$. Note that k_0 exists by Assumptions (1.3) and (1.10). By Proposition 1.7 the Hopf index H_n of ODE $(1.2)_n$ satisfies

$$0 \neq H = \lim_{\nu \to \infty} H_\nu = H_n \quad \text{for } n \geqq n_0.$$

The approximation leaves the set of stationary solutions unchanged, no eigenvalues zero occur in B_k for $n \geqq n_0$. From the proof of [15, Thm. 4.7] we conclude that there exist continua $Z_k^n \subset I \times X$ of centers and nonstationary periodic solutions of $(1.1)_n$ with the following properties:

(3.5) $Z_k^n \subset B_k$ and the prime periods on Z_k^n are $\leqq 2k$,

(3.6) Z_k^n contains a center z^n of $(1.1)_n$, independently of k,

(3.7) Z_k^n is compact and connected,

(3.8) if $Z_k^n \cap \partial B_k$ is empty, then there is a continuum $Z_k'^n \subset Z_k^n$ such that

(i) any element of $Z_k'^n$ has some virtual period $\geqq c_0$ where c_0 was defined in Lemma 2.4, and

(ii) any interval $[\tau, 2\tau] \subset [c_0, 2k]$ contains a virtual period of an element of $Z_k'^n$.

The Z_k^n, $Z_k'^n$ correspond to the continua Z_k, Z_k' which were constructed in [15]: any element $(\lambda,y)\in I\times\mathbb{R}^m$ of Z_k is reinterpreted as an element $(\lambda,y(t))\in I\times X$, attaching the orbit to it. Properties (3.5), (3.6), (3.8) are read off from properties [15, (4.9)–(4.12)] of the corresponding snakes. The assumption [15, (1.2)] on C^4 differentiability of g^n may be dropped for ODEs, using smoothed generic approximations. Assumption [15, (4.5)] on analyticity was not used in the construction of Z_k, Z_k'. Thus existence of Z_k^n is established.

Step 2: *construction of* $Z_k=$"lim" Z_k^n. We define

$$Z_k := \bigcap_{\nu\geqq n_0}\left(\overline{\bigcup_{n\geqq\nu} Z_k^n}\right)$$

$$= \left\{ z \mid z=\lim z_{n_j} \text{ for some sequences } n_j\to\infty,\ z_{n_j}\in Z_k^{n_j}\right\}.$$

Obviously, $Z_k\subset B_k$ is closed. By (3.5) and compactness Lemma 1.1, the set Z_k consists of periodic solutions of IE (1.1) with period $\leqq 2k$, and Z_k is compact. Restricting the whole approximation process to a subsequence of $n\in\mathbb{N}$ from now on, we may also assume that the $z^n\in Z_k^n$ from (3.6) converge to

$$z_0 := \lim_{n\to\infty} z^n\in C,$$

which implies $z_0\in Z_k$ for all $k\geqq k_0$.

Moreover Z_k is connected: otherwise we would have a partition $Z_k=Z'\cup Z''$ into disjoint, nonempty, compact subsets with $z_0\in Z''$ and $\alpha := \frac{1}{2}\operatorname{dist}(Z',Z'')>0$. By definition of Z_k there exists ν_0 such that for all $\nu>\nu_0$ we have

$$\alpha>\operatorname{dist}\left(Z_k^\nu,Z_k\right)=\operatorname{dist}\left(Z_k,Z'\cup Z''\right).$$

Now $z_0\in Z''$ implies that $z^\nu\in Z_k^\nu\cap N_\alpha(Z'')$ for large ν and the open α-neighborhood $N_\alpha(Z'')$. Hence $Z_k^\nu\cap N_\alpha(Z')=\varnothing$, because Z_k^ν is connected. This is a contradiction to $Z'\neq\varnothing$, therefore Z_k is connected.

Step 3: *centers in* Z_k. We claim that Z_k is not contained in the set of centers C. Analyticity is used only in this step. We consider two cases.

CASE 1. There exists a sequence $n_j\to\infty$ (which may depend on k) such that $Z_k^{n_j}\cap\partial B_k$ is nonempty.

Then $Z_k\cap\partial B_k$ is nonempty and Z_k is not contained in $C\subset B_k-\partial B_k$, for $k\geqq k_0$.

CASE 2. For all $n\geqq n_0(k)$, $Z_k^n\cap\partial B_k$ is empty.

Then (3.8), (i) and (ii) are satisfied, i.e. the virtual periods are large on a continuum $Z_k'^n$. Analogously to Z_k, we construct a continuum Z_k' of solutions of (1.1) with large virtual periods: by (3.8), (ii) there exist z_k^n in $Z_k'^n$ with some virtual period in $[c_0,2c_0]$. Passing to subsequences we may assume that

$$z_k := \lim_{n\to\infty} z_k^n$$

exists. We define

$$Z_k' := \bigcap_{\nu\geqq n_0(k)}\overline{\bigcup_{n\geqq\nu} Z_k'^n}$$

and z_k is an element of the compact connected set Z_k'. Moreover the virtual period Proposition 2.2 implies that (3.8), (i)–(ii) remains valid if we replace $Z_k'^n$ by $Z_k'\subset Z_k$.

We show that $Z_k' \not\subset C$, by contradiction. If $Z_k' \subset C$ then Z_k' consists of just one point z_k because, by Lemma 2.4, the set $J(c_0)$ defined there is dense in C and has empty intersection with the connected set Z_k', by (3.8)(i). But then, by (3.8)(ii), the point z_k has at least 2^j different virtual periods for $k \geqq k_0$ if we choose

$$k_0 > 2^{2^j} c_0.$$

This is a contradiction to Lemma 2.3, if we choose $2j$ to be a uniform bound on C for the number of characteristic roots of χ on the imaginary axis: then $z_k \in C$ has less than 2^j distinct virtual periods.

Step 4: *finale for* Z. We define the set Z of centers and nonstationary periodic solutions of IE (1.1) as

$$Z := \bigcup_{k \geqq k_0} Z_k.$$

By Step 2, the connected sets Z_k have the center z_0 in common, hence Z is connected. By Step 3, Z is not contained in C. Again by Step 3, Z_k contains a *nonstationary* periodic solution z which lies on ∂B_k (Case 1), or has a virtual period between $2^{-2^j-1}k$ and $2k$ (Case 2). Therefore Z is global, i.e. Z satisfies (3.2) or (3.3), and the proof is complete. □

4. Applications. We indicate applications of our abstract global bifurcation Theorem 3.1 to two problems arising in mathematical biology. We compute the Hopf index on suitably chosen intervals I of the parameter λ. In the first problem we have to adapt the solution space X to the biological situation. Our results are stated as corollaries to Theorem 3.1.

The first example is drawn from Diekmann, Montijn [11] and Gripenberg [16], [17] and was the original motivation of our work. The scalar equation, arising in epidemiology, reads [11]

$$x(t) = \lambda \cdot \left(1 - \int_{t-1}^{t} x(s)\,ds\right) \cdot \int_{-\infty}^{t} b(t-s)x(s)\,ds; \tag{4.1}$$

and we assume $b \geqq 0$ has integral 1 and is in A_γ for some $\gamma > 0$. The parameter λ relates to the total population size, $x(s) \geqq 0$ counts the individuals which are newly infected at time s, and $b(\tau)$ measures the infective force of an individual, which was infected τ units of time ago, on susceptible individuals (see [11] for more details and references). Let

$$\xi_1(t) := \int_{t-1}^{t} x(s)\,ds,$$

$$\xi_2(t) := \int_{-\infty}^{t} b(t-s)x(s)\,ds,$$

$$g(\xi_1, \xi_2) := (1-\xi_1)\xi_2.$$

Use (4.1) to replace $x = \lambda g(\xi_1, \xi_2)$ in the definition of ξ_1, ξ_2 to obtain the equivalent system

$$\xi = \lambda a * f(\xi), \quad \text{where} \tag{4.2}$$

$$\xi = \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix}, \quad f = \begin{pmatrix} g \\ g \end{pmatrix}, \quad a = \begin{pmatrix} 1_{[0,1]} & 0 \\ 0 & b \end{pmatrix}.$$

Thus (4.1) is put into the general form (1.1) discussed before. As a solution space for (4.1) we consider the subset

$$x \in X := BC^0(\mathbb{R}, \mathbb{R}^+)$$

of the Banach space $BC^0(\mathbb{R}, \mathbb{R})$ i.e. we admit positive solutions x only.

Let us analyze the stationary solutions and characteristic equation of (4.1). In X, the only stationary solution is

$$x^0(\lambda) = 1 - 1/\lambda, \qquad \lambda > 1$$

(we exclude the branch $x(\lambda) \equiv 0 \notin X$), with corresponding characteristic function

$$\chi_\lambda(\mu) = 1 + (\lambda - 1)\mu^{-1} \cdot (1 - \exp(-\mu)) - \hat{b}(\mu).$$

We discuss the limits $\lambda \searrow 1$ and $\lambda \to \infty$ in order to compute the Hopf index along $x^0(\lambda)$. The limit $\lambda \searrow 1$ is regular, $\mu = 0$ is the only zero of $\chi_1(\mu)$ with $\operatorname{Re}\mu \geqq 0$. This simple zero continues to $\mu = \mu(\lambda)$ near $\lambda = 1$ and from

$$\mu'(1) = \left(\hat{b}'(0)\right)^{-1} = -\left(\int_0^\infty s b(s)\, ds\right)^{-1} < 0$$

we conclude that

$$\lim_{\lambda \searrow 1} E^0(\lambda) = 0 \tag{4.3}$$

(the "unstable dimensions" $E^j(\lambda)$ were introduced just before Definition 1.6). The limit $\lambda \to \infty$ is also easy [11]: with $\beta := 1/(\lambda - 1)$, $\chi = 0$ is equivalent to

$$\eta(\beta, \mu) := \beta(1 - \hat{b}(\mu)) + \mu^{-1}(1 - \exp(-\mu)) = 0 \tag{4.4}$$

where $\eta : \mathbb{R} \times \mathbb{C} \to \mathbb{C}$ is analytic. Note that $\eta(0, \mu_k) = 0$ at the simple zeros $\mu_k = k \cdot 2\pi i$, $k \in \mathbb{Z} \setminus \{0\}$. By the implicit function theorem, these zeros have continuations $\mu_k(\beta)$ for small $|\beta|$ with derivative

$$\mu_k'(0) = -\left(D_\mu \eta(0, \mu_k)\right)^{-1} D_\beta \eta(0, \mu_k) = -\mu_k^{-1}(1 - \hat{b}(\mu_k)).$$

In the notation of [11], $b_k := 2 \operatorname{Im} \hat{b}(\mu_k)$, we obtain

$$\operatorname{Re}\mu_k'(0) = \frac{b_k}{4k\pi}.$$

This implies that

$$E^0(\lambda_0) > 0 \quad \text{for large } \lambda_0 > 1, \tag{4.5}$$

if at least one of the numbers b_k is positive.

We check the assumptions of our Main Theorem 3.1 for $\lambda \in I := [1 + \varepsilon, \lambda_0]$ and $\varepsilon > 0$ sufficiently small. We get around the boundedness assumption on f as in our remarks on Theorem 3.1. Obviously, $b \in A_\gamma$ implies $a \in A_\gamma$. Assumption (1.9) is satisfied: we obtain a stationary branch $x^0(\lambda)$ with $\chi_\lambda(0) = \lambda - 1 \neq 0$. The set C of centers is discrete. To see that we write $\mu = i\tau$ and solve equation (4.4) for β, obtaining

$$\beta = \Phi(\tau)$$

with $\Phi(\tau) := -(1-\exp(-i\tau))\cdot(i\tau(1-\hat{b}(i\tau)))^{-1}$. The curve Φ is analytic for $\tau>0$ and $\operatorname{Im}\Phi \not\equiv 0$ because $\Phi(2\pi k)=0$, $\operatorname{Im}\Phi'(2\pi k)\neq 0$ provided that $b_k\neq 0$. Together with $\Phi(\tau)\to 0$ for $\tau\to\infty$ this proves C is discrete. Hence we may pick λ_0 large enough such that (4.5) is valid and no centers occur on ∂I, i.e. Assumption (1.10) holds. By Definition 1.6 of the Hopf index, (4.3) and (4.5) imply

$$H=\frac{1}{2}(-1)^{E^0(1+\varepsilon)}\cdot\left(E^0(\lambda_0)-E^0(1+\varepsilon)\right)=\frac{1}{2}E^0(\lambda_0)>0,$$

i.e. Assumption (3.1) holds. Analyticity of $\chi_\lambda(\mu)$ (cf. (2.7)) is trivial. Applying Theorem 3.1 we obtain the following.

COROLLARY 4.1. *Let the kernel b in equation* (4.1) *be nonnegative with integral normalized to one, $b\in A_\gamma$ for some positive γ, and*

$$b_k := 2\int_0^\infty b(s)\,\sin(2\pi k s)\,ds>0$$

for some $k\in\mathbb{N}$.

Then one of the following holds:

(4.6) *For any large enough λ_0 which does not occur in the discrete set C of centers there exists a nonstationary periodic solution $x\in X=BC^0(\mathbb{R},\mathbb{R}^+)$ of equation* (4.1); *or*

(4.7) *there exists a $\lambda_0>1$ and a continuum $Z\subset(1,\lambda_0)\times X$ consisting of centers and nonstationary periodic solutions of* (4.1) *such that Z is global, i.e. Z contains nonstationary periodic solutions (λ,x) with*
- *arbitrarily large norms of x, or*
- *arbitrarily large virtual periods.*

Proof. Our previous remarks prove that a global continuum $Z_\varepsilon\subset[1+\varepsilon,\lambda_0]\times X$ exists, if we can exclude the possibility that Z_ε terminates at a periodic orbit in $[1+\varepsilon,\lambda_0]\times\partial X$. From Gripenberg [17] it is immediate that 0 is the only periodic orbit in ∂X.

If $(\lambda,0)$ is in $\overline{Z}_\varepsilon$ with virtual periods staying bounded on Z, then $(\lambda,0)$ is a center, $\lambda\neq 1$ and there is a periodic solution $y\geqq 0$, $|y|=1$ of the linearized equation

$$y=\lambda b * y.$$

Taking averages over a period, we obtain a contradiction to $\lambda\neq 1$, $\int b=1$. Thus $(\lambda,0)\notin\overline{Z}$ and Theorem 3.1 applies on $[1+\varepsilon,\lambda_0]\times X$.

Letting $\varepsilon\to 0$, we can in fact obtain a global continuum $Z=$"lim" Z_ε in $[1,\lambda_0]\times X$ as was indicated in our remarks following Theorem 3.1; cf. also Step 2 in the proof of Theorem 3.1.

Note that (4.1) cannot have nontrivial periodic orbits at $\lambda=1$, to complete the proof of the corollary. □

Under the stronger but epidemiologically reasonable assumption that there is a constant C such that $b\leqq C\cdot 1_{[0,1]}$ it is easy to see that all periodic solutions in X have norm at most $\lambda\cdot C$. Then Corollary 4.1 holds even if we drop the alternative "Z is unbounded" in (4.7).

Our second example concerns a model for circular neural nets, cf. an der Heiden [22]. The system reads

$$x_j=\lambda h_j * f_j(x_{j-1}),\qquad j=1,\cdots,m \tag{4.8}$$

with $x_0 := x_m$. We assume

(4.9) the kernels $h_j \geqq 0$ have integral normalized to 1, and $h_j \in A_\gamma$ for some $\gamma > 0$ and all j; the C^1-nonlinearities f_j satisfy $f_j(\xi) \neq 0$ for $\xi \neq 0$, $f_j(0) = 0$ and $c := \Pi_j f_j'(0) < 0$.

Systems (4.8) with $\lambda > 0 > c$ are called "repressible" in [22]. We give a general discussion of the Hopf index H first and specify kernels h_j later. Our result shows how *delays* in the kernels cause global Hopf bifurcation.

This time we work directly in the setting of equation (1.1), $I = [0, \lambda_0]$. The only stationary solution of system (4.8) is the zero solution (by Assumption 4.9), and the characteristic equation reads

$$\chi_\lambda(\mu) = 1 - \lambda^m \cdot c \prod_j \hat{h}_j(\mu) = 0.$$

Obviously, $\chi_\lambda(0) = 1 - \lambda^m c \neq 0$.

For $\lambda = 0$ we trivially obtain $\chi(\mu) \equiv 1$, hence

$$E(\lambda = 0) = 0. \tag{4.10}$$

Obviously, nonstationary periodic solutions do not exist for $\lambda = 0$.

Analysis for $\lambda = \lambda_0 > 0$ is much more involved. Introducing $\beta := \lambda^{-m}$, $\mu = i\tau$ and $\Phi(\tau) := c \cdot \Pi \hat{h}_j(i\tau)$, the characteristic equation reads

$$\Phi(\tau) = \beta. \tag{4.11}$$

As before, Φ is analytic. Therefore $\Phi(0) = c$, $\operatorname{Im} \Phi'(0) \neq 0$ implies that $\operatorname{Im} \Phi \not\equiv 0$ and the set C of centers is discrete, by real analyticity of $\tau \to \operatorname{Im} \Phi(\tau)$. However, C may be unbounded.

In our first example, we have discussed the case that Φ has a real zero. This time, let us assume $\Phi(\tau) \neq 0$ for all real τ. To compute

$$E(\lambda_0)$$

we may perturb Φ slightly, such that $\Phi'(\tau) \neq 0$ whenever $\Phi(\tau) \in \mathbb{R}^+$. In that case, all imaginary characteristic roots $\mu = i\tau$ are simple and (4.11) may be solved for $\mu = \mu(\beta)$, locally, with

$$\operatorname{sgn} \operatorname{Re} \mu'(\beta) = -\operatorname{sgn} \operatorname{Im} \tau'(\beta) = \operatorname{sgn} \operatorname{Im} \Phi'(\tau). \tag{4.12}$$

This equality tells us how $E(\lambda)$ relates to the winding number $n(\Phi, 0)$ around 0 of the complex valued curve $\Phi : \mathbb{R} \to \mathbb{C} \setminus \{0\}$. For $(\lambda_0, 0) \notin C$ and $\beta_0 := \lambda_0^{-m}$ let $\tau_0 > 0$ be maximal (Riemann–Lebesgue lemma) such that $\Phi(\tau_0) \in [\beta_0, \infty)$. Then all imaginary characteristic roots $\mu = i\tau$ on $C \cap ([0, \lambda_0] \times \{0\})$ satisfy $|\tau| \leqq \tau_0$ and contribute to the left-hand side of (4.12) for some $\tau \in [-\tau_0, \tau_0]$, adding up to $E(\lambda_0)$. The right-hand sides add up almost to the winding number $n(\Phi([-\tau_0, \tau_0]), 0)$ with a possible difference of at most one. Thus we have proved

$$|E(\lambda_0) - n(\Phi([-\tau_0, \tau_0]), 0)| \leqq 1. \tag{4.13}$$

We apply our Main Theorem 3.1 with $\lambda \in I = [0, \lambda_0]$ just as with our first example. By (4.10), we know that on $[0, \lambda_0] \times X$

$$H = \frac{1}{2} E(\lambda_0),$$

and (4.13) helps us to compute this quantity.

Before we specify particular kernels we summarize our result as follows.

COROLLARY 4.2. *Let Assumption* (4.9) *be satisfied for the repressible system* (4.8). *In addition suppose that for* $\lambda=\lambda_0>0$ *the winding number* $n(\Phi,0)$ *satisfies*

$$|n(\Phi([-\tau_0,\tau_0]),0)|>1. \tag{4.14}$$

Then the conclusion of Theorem 3.1 *holds for system* (4.8) *on* $[0,\lambda_0]\times X$. *Obviously,* (3.4) $\lambda=\alpha=0$ *is impossible, i.e. nonstationary periodic orbits do not exist at* $\lambda=0$.

As a simple example we specify the following *delaying* kernels

$$h_j(s)=\begin{cases}0 & \text{for } 0\leqq s<\zeta_j,\\ b_j(s-\zeta_j) & \text{for } \zeta_j\leqq s<\infty,\end{cases} \tag{4.15}$$

$$b_j(s):=p_j(s)\exp(-\gamma_j s),$$

with $\gamma_j>0$, $\zeta_j\geqq 0$, $\Sigma\zeta_j>0$ and nonzero polynomials $p_j\geqq 0$ on $[0,\infty)$. We check Assumption (4.14) on the winding number of Φ. We compute

$$\begin{aligned}\Phi(\tau)&=\prod_j f_j'(0)\hat{h}_j(i\tau)\\ &=c\cdot\prod_j\hat{b}_j(i\tau)\cdot\exp\Big(-i\tau\sum_j\zeta_j\Big)\\ &=c\cdot\prod_j\hat{p}_j(\gamma_j+i\tau)\cdot\exp\Big(-i\tau\sum_j\zeta_j\Big)\\ &=c\cdot\prod_j q_j\big(1/(\gamma_j+i\tau)\big)\cdot\exp\Big(-i\tau\sum_j\zeta_j\Big),\end{aligned}$$

where q_j are polynomials related to p_j. Thus

$$\Phi(\tau)=cq(\tau)\exp\Big(-i\tau\sum_j\zeta_j\Big),$$

with $\arg(q(\tau))$ being uniformly bounded for $\tau\in\mathbb{R}$ (we take $\arg(q(\tau))$ continuous as a function of τ). By equation (4.11), the set C of centers is therefore unbounded and

$$\lim_{\tau_0\to\infty} n(\Phi([-\tau_0,\tau_0]),0)=-\infty.$$

Hence Corollary 4.2 applies for all sufficiently large $\lambda_0>0$. This result is not affected by a possible finite number of real zeros of Φ.

Following our previous analysis we could treat other kernels (e.g. step functions) as well. Also, small self-inhibition or self-excitation in (4.8) does not change our results.

Finally, we note that Gripenberg's example [16], [17] can be treated along the lines of our second example. With regard to $\Phi(\tau)$, (4.1) is rather special because it yields infinitely many zeros of Φ.

Acknowledgments. The author is indebted to W. Jäger, K. Schumacher and H. Thieme for teaching him some integral equations with patience and enthusiasm, to O. Diekmann who initiated this research by his questions, and to the referees for their constructive criticism.

Note added in proof. Above we announce a result on global Hopf bifurcation if Assumption (1.9) is violated. This accounts for degenerate stationary solutions and their bifurcations. The corresponding ODE theory is developed in B. Fiedler, *Global Hopf bifurcation of two parameter flows*, Arch. Rational Mech. Anal., to appear. By ODE-approximation, this theory carries over to Volterra integral equations as considered in the present paper.

REFERENCES

[0] J. C. ALEXANDER AND J. A. YORKE, *Global bifurcations of periodic orbits*, Amer. J. Math., 100 (1978), pp. 263–292.

[1] K. T. ALLIGOOD, J. MALLET-PARET AND J. A. YORKE, *Families of periodic orbits: local continuability does not imply global continuability*, J. Differential Geom., 16 (1981), pp. 483–492.

[2] _____, *An index for the global continuation of relatively isolated sets of periodic orbits*, in Geometric Dynamics, Palis Jr., ed., Rio de Janeiro, 1981, Lecture Notes in Mathematics 1007, Springer-Verlag, New York, Berlin, Heidelberg, pp. 1–21.

[3] K. T. ALLIGOOD AND J. A. YORKE, *Families of periodic orbits: virtual periods and global continuability*, J. Differential Equations, 55 (1984), pp. 59–71.

[4] W. ALT, *Some periodicity criteria for functional differential equations*, Manuscripta Math., 23 (1978), pp. 295–318.

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

[6] S.-N. CHOW AND J. MALLET-PARET, *The Fuller index and global Hopf bifurcation*, J. Differential Equations, 29 (1978), pp. 66–85.

[7] S.-N. CHOW, J. MALLET-PARET AND J. A. YORKE, *A periodic orbit index which is a bifurcation invariant*, in Geometric Dynamics, Palis Jr., ed., Rio de Janeiro, 1981, Lecture Notes in Mathematics 1007, Springer-Verlag, New York, Berlin, Heidelberg, pp. 109–131.

[8] J. M. CUSHING, *Bifurcation of periodic solutions of integrodifferential systems with applications to time delay models in population dynamics*, SIAM J. Appl. Math., 33 (1977), pp. 640–654.

[9] _____, *Integrodifferential equations and delay models in population dynamics*, Lecture Notes in Biomathematics 20, Springer-Verlag, New York, Berlin, Heidelberg, 1977.

[10] _____, *Nontrivial periodic solutions of some Volterra integral equations*, Helsinki Symposium on Integral Equations 1978, Londen and Staffans, eds., Springer-Verlag, New York, Berlin, Heidelberg, pp. 50–66.

[11] O. DIEKMANN AND R. MONTIJN, *Prelude to Hopf bifurcation in an epidemic model: analysis of a characteristic equation associated with a nonlinear Volterra integral equation*, J. Math. Biol., 14 (1982), pp. 117–127.

[12] O. DIEKMANN AND S. A. VAN GILS, *Invariant manifolds for Volterra integral equations of convolution type*, J. Differential Equations, 54 (1984), pp. 139–180.

[13] G. EISENACK AND C. FENSKE, *Fixpunkttheorie*, BI, Mannheim, 1978.

[14] B. FIEDLER, *Global Hopf bifurcation in porous catalysts*, Proceedings Equadiff 82, H. W. Knobloch and K. Schmitt, eds., Würzburg, 1982, Lecture Notes in Mathematics 1017, Springer-Verlag, New York, Berlin, Heidelberg, pp. 177–183.

[15] _____, *An index for global Hopf bifurcation in parabolic systems*, Crelle J. Reine Angew. Math., 359 (1985), pp. 1–36.

[16] G. GRIPENBERG, *Periodic solutions of an epidemic model*, J. Math. Biol., 10 (1980), pp. 271–280.

[17] _____, *On some epidemic models*, Quart. Appl. Math., 39 (1981), pp. 317–327.

[18] _____, *Stability of periodic solutions of some integral equations*, Crelle J. Reine Angew. Math., 331 (1982), pp. 16–31.

[19] K. P. HADELER AND J. TOMIUK, *Periodic solutions of difference-differential equations*, Arch. Rational Mech. Anal., 65 (1977), pp. 87–95.

[20] J. K. HALE, *Theory of Functional Differential Equations*, Springer-Verlag, New York, Berlin, Heidelberg, 1977.

[21] J. K. HALE AND J. C. F. DE OLIVEIRA, *Hopf bifurcation for functional equations*, J. Math. Anal. Appl., 74 (1980), pp. 41–59.

[22] U. AN DER HEIDEN, *Analysis of Neural Networks*, Lecture Notes in Biomathematics 35, Springer-Verlag, New York, Berlin, Heidelberg, 1980.

[23] H. W. HETHCOTE, H. W. STECH, P. VAN DEN DRIESSCHE, *Periodicity and stability in epidemic models: a survey*, in Differential Equations and Applications in Ecology, Epidemics and Population Problems, S. N. Busenberg and K. L. Cooke, eds., Academic Press, New York, 1981.

[24] J. IZE, *Periodic solutions of nonlinear parabolic equations*, Comm. Partial Differential Equations, 4 (1979), pp. 1299–1387.

[25] _____, *Obstruction theory and multiparameter Hopf bifurcation*, preprint.

[26] N. D. KAZARINOFF, Y.-H. WAN AND P. VAN DEN DRIESSCHE, *Hopf bifurcation and stability of differential-difference and integro-differential equations*, J. Inst. Math. Appl., 21 (1978), pp. 461–477.

[27] J. MALLET-PARET AND J. A. YORKE, *Snakes: oriented families of periodic orbits, their sources, sinks and continuation*, J. Differential Equations, 43 (1982), pp. 419–450.

[28] R. K. MILLER AND G. R. SELL, *Volterra integral equations and topological dynamics*, AMS memoir, 102, 1970.

[29] R. D. NUSSBAUM, *The radius of the essential spectrum*, Duke Math. J., 37 (1970), pp. 473–478.

[30] _____, *Periodic solutions of nonlinear autonomous functional differential equations*, in Functional Differential Equations and Approximation of Fixed Points, H.-O. Peitgen and H.-O. Walther, eds., Bonn, 1978, Lecture Notes in Mathematics 730, Springer-Verlag, New York, Berlin, Heidelberg, pp. 283–325.

[31] _____, *A global bifurcation theorem with applications to functional differential equations*, J. Funct. Anal., 19 (1975), pp. 319–338.

[32] K. SCHUMACHER, *Hopf bifurcation with constraints*, Nonlinear Anal. Theory Meth. Appl., 7 (1983), pp. 1389–1409.

[33] H. SMITH, *Hopf bifurcation in a system of functional equations modeling the spread of an infectious disease*, SIAM J. Appl. Math., 43 (1983), pp. 370–385.

[34] H.-O. WALTHER, *Über Ejektivität und periodische Lösungen bei autonomen Funktionaldifferentialgleichungen mit verteilter Verzögerung*, Habilitationsschrift, München, 1977.

SIAM J. MATH. ANAL.
Vol. 17, No. 4, July 1986

012

KINETIC EQUATIONS WITH REFLECTING BOUNDARY CONDITIONS*

C. V. M. VAN DER MEE† AND V. PROTOPOPESCU‡

Abstract. A general abstract model of time-independent kinetic equations on the half-line is presented. The existence and uniqueness of the solution is proved under specified incoming flux and nonmultiplying boundary reflection processes. An iterative method is formulated for computing in principle the solution by using the solution of the analogous problem without reflection. In many concrete cases (e.g. neutron transport, BGK model in rarefied gas dynamics, etc.) the available explicit expression for the latter provides the actual solution of the general problem. Possible generalizations and open problems are briefly discussed.

Key words. kinetic theory, transport equation, reflection

AMS(MOS) subject classifications. Primary 82A70; secondary 45A25

1. Introduction. In recent years substantial progress has been reported on the existence and uniqueness theory for the solution of boundary value problems of the type

$$T\psi'(x) = -A\psi(x), \qquad x \in \mathbb{R}_+, \tag{1.1}$$

$$Q_+\psi(0) = \mathscr{R}JQ_-\psi(0) + \phi_+, \tag{1.2}$$

$$\|\psi(x)\| = O(1) \qquad (x \to \infty), \tag{1.3}$$

where T is an injective self-adjoint operator, Q_+ and Q_- the orthogonal projections onto the maximal positive and negative T-invariant subspaces and A a positive self-adjoint (bounded or unbounded) Fredholm operator. The operators $\mathscr{R}$ and J as well as the precise meaning of the norm in (1.3) will be specified later. This boundary value problem models a variety of time-independent transport phenomena in semi-infinite media with boundary conditions appropriate to incoming flux specification and, if $\mathscr{R}$ is nonzero, to a (partial) reflection at the boundary. In most instances, however, it has been assumed that $\mathscr{R}=0$ (absence of reflection), and in this case the solution ψ, whenever unique, is represented in the form

$$\psi(x) = e^{-xT^{-1}A}E\phi_+, \qquad x \in \mathbb{R}_+. \tag{1.4}$$

In this direction we note the important contributions of Hangelbroek [14], Lekkerkerker [16], Beals [1, 2], van der Mee [19] and Greenberg et al. [13]. Only recently such a theory has been developed with full account of boundary reflection processes ($\mathrm{R} \equiv 0$). Namely, Beals and Protopopescu [3], [4] obtained an existence and uniqueness theory

*Received by the editors May 29, 1984, and in revised form November 14, 1984. This research was supported in part by the U.S. Department of Energy under grant DE-AS05 80ER10711-1 and by the National Science Foundation under grant DMS-8312451.

†Department of Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061. Present address, Department of Mathematics, Texas Tech University, Lubbock, Texas 79409.

‡Department of Chemistry, Boston University, Boston, Massachusetts 02215.

for the Fokker–Planck equation

$$(1.5)\quad v\frac{\partial\psi}{\partial x}(x,v)=\frac{\partial^2\psi}{\partial v^2}(x,v)-v\frac{\partial\psi}{\partial x}(x,v),\qquad x\in\mathbb{R}_+,\quad v\in\mathbb{R},$$

$$(1.6)\quad \psi(0,v)=\alpha\psi(0,-v)+\beta\int_{-\infty}^{0}\sigma(v'\to v)\psi(0,v')\,dv'+\phi_+(v),\qquad v\in\mathbb{R}_+,$$

$$(1.7)\quad \lim_{x\to\infty}\{\psi(x,v)/x\}=b.$$

For general σ's the existence proof required a condition which will turn out to be automatically satisfied. Maslova [17], [18] published related results on the linearized Boltzmann equation with a sufficiently regular intermolecular potential. Greenberg and van der Mee [12] formulated a related radiative transfer problem in an abstract setting, but their result deals with (1.1) on a finite interval. The abstract approach was recently followed by van der Mee [20], who announced some results on this problem.

The paper is organized as follows. For the reader's convenience in §2 we provide a brief but fairly complete review of the existence and uniqueness theory for the solution of (1.1)–(1.3). Section 3 contains the procedure of *computing* the solution of the problem with reflection from the solution of the same problem without reflection ($\mathscr{R}=0$). This iterative scheme can be implemented whenever the albedo operator E in (1.4) is known explicitly, which is actually the case for a large class of problems. Namely, if A is a compact perturbation of the identity, an expression for E in terms of generalized Chandrasekhar **H**-functions [9] has been given by van der Mee [21], thereby generalizing a plethora of results obtained before for specific models (neutron transport, radiative transfer, BGK models, etc.). In some cases the iterative procedure can also be derived from a half-range completeness result involving **H**-functions (cf. [7], for instance). At present, no explicit representation is known for the albedo operator E in the case of the Fokker–Planck model (1.5)–(1.7) and more general Sturm–Liouville problems.

2. Existence and uniqueness theory. In the present section we provide a brief but fairly complete review of the existence and uniqueness theory for the solution of (1.1)–(1.3). In order to explain the later introduction of a number of concepts, we present the overall flavor of the theory by first considering T bounded and A strictly positive and neglecting reflection processes, which is relevant to radiative transfer in absorbing atmospheres [9] and neutron transport in submultiplying reactors [7]. Using semigroup theory, one may naturally write solutions to (1.1) in the form

$$\psi(x)=e^{-xT^{-1}A}\psi(0),\qquad 0\leqq x<\infty,$$

where $\psi(0)$ must be chosen in the subspace corresponding to the nonnegative part of the spectrum of the evolution operator $T^{-1}A$ in order that the above semigroup expression makes sense and condition (1.3) is fulfilled. On fitting the boundary condition (1.2) where $\mathscr{R}=0$, one must require $Q_+\psi(0)=\varphi_+$. If one would formulate an analogous boundary value problem for $x\in(+\infty,0)$, one should require that $\psi(0)$ be chosen in the subspace corresponding to the nonpositive part of the spectrum of $T^{-1}A$ and $Q_-\psi(0)=\varphi_-$. In a natural way one may thus express the unique solvability of both half-space problems, for $x\in(0,\infty)$ and for $x\in(-\infty,0)$, in terms of the invertibility of the operator V, which maps the nonnegative (resp. nonpositive) spectral subspace of $T^{-1}A$ into the ranges of the projections $Q_\pm$ onto forward (resp. backward) "fluxes". As a matter of fact, $V\psi(0)=\varphi_+$, $E=V^{-1}$ is called the albedo operator and formula (1.4) arises as the obvious result. Below we shall review the existence and uniqueness theory

along the above set up (which originates from Hangelbroek [14]) in some detail, since the unboundedness of T, the appearance of a nonzero null space of A and the absence of compactness assumptions on A cause technical difficulties and necessitate the introduction of some novel notions.

Let us now drop the above restrictions on T and A. Let T be an injective self-adjoint operator and A a positive self-adjoint operator with closed range $\mathrm{Ran}A$ and null space $\mathrm{Ker}A$ of finite dimension, both defined on the complex Hilbert space H. For the sake of convenience we assume A to be bounded, but at the end of this work we shall discuss how to remove this restriction. We then define the zero root subspace

$$Z_0=\{h\in H/\exists n\in\mathbb{N}:(T^{-1}A)^n h=0\}, \tag{2.1}$$

and assume $Z_0\subset D(T)$. It can then be proved (cf. [13]; the result there extends to unbounded T) that Z_0 has a finite dimension and $Z_0=\mathrm{Ker}(T^{-1}A)^2$. Here we also assume that Z_0 is nondegenerate in the following sense:

$$\{h\in Z_0/(Th,k)=0 \text{ for all } k\in Z_0\}=\{0\}.$$

In fact, this assumption is automatically satisfied. (If T is bounded, see [13]).

PROPOSITION 2.1. *We have the following decompositions*:

$$Z_0\oplus(T[Z_0])^{\perp}=H, \tag{2.2}$$

$$T[Z_0]\oplus Z_0^{\perp}=H, \tag{2.3}$$

$$Z_0^{\perp}=\overline{T\{(T[Z_0])^{\perp}\}}=A\{(T[Z_0])^{\perp}\}. \tag{2.4}$$

Moreover, Z_0 and $(T[Z_0])^{\perp}$ are $T^{-1}A$-invariant subspaces and there exists a unique operator S on $(T[Z_0])^{\perp}$ such that

$$T^{-1}A=(T^{-1}A|_{Z_0})\oplus S^{-1}. \tag{2.5}$$

The operator S is self-adjoint with respect to the positive definite inner product on $(T[Z_0])^{\perp}$ given by

$$(h,k)_A=(Ah,k). \tag{2.6}$$

For isotropic neutron transport in a conservative medium, where Z_0 can be constructed explicitly, this Proposition 2.1 is due to Lekkerkerker [16]. It later appeared in more abstract form in [19], [13], [2].

Let us introduce H_T as the Hilbert space obtained by completing $D(T)$ with respect to the inner product

$$(h,k)_T=(|T|h,k).$$

Let us assume that there exists a unitary and self-adjoint operator J on H, which leaves invariant $D(T)$ and satisfies

$$TJ=-JT, \qquad AJ=JA.$$

Then J extends from $D(T)$ to a unitary and self-adjoint operator on H_T, as also do the orthogonal projections Q_+ and Q_- of H onto the maximal positive and negative T-invariant subspaces, respectively. We shall require the *reflection operator* to be a bounded operator on $Q_+[H]$, which leaves invariant $D(T)$, and satisfies the identity

$$T\mathscr{R}=\mathscr{R}^{\dagger}T \tag{2.7}$$

for some bounded operator $\mathscr{R}^+$ on H. Putting

$$\mathscr{R}h \overset{\text{def}}{=} J\mathscr{R}Jh, \qquad h \in Q_-[H],$$

we extend $\mathscr{R}$ to a bounded operator on H, which automatically commutes with J, leaves invariant $D(T)$ and extends from $D(T)$ to a bounded operator on H_T. In order that the operator $\mathscr{R}$ models a nonmultiplying reflection process at the boundary, we also assume that $\mathscr{R}$ extends to a contraction on H_T:

$$(|T|\mathscr{R}h, \mathscr{R}h) \leqq (|T|h, h), \qquad h \in D(T).$$

For later use we include the following result, derived by Beals [1] for injective and certain noninjective A, and generalized by Greenberg et al. [13].

THEOREM 2.2. *For every $\phi_+ \in Q_+[H_T]$ there exists at least one continuous function $\psi\colon [0,\infty) \to H_T$, which is continuously differentiable on $(0,\infty)$ and satisfies the equations*

$$T\psi'(x) = -A\psi(x), \qquad x \in \mathbb{R}_+, \tag{2.8}$$

$$Q_+\psi(0) = \phi_+, \tag{2.9}$$

$$\|\psi(x)\|_T = O(1) \qquad (x \to \infty). \tag{2.10}$$

The number of linearly independent solutions of the homogeneous $(\phi_+ = 0)$ problem coincides with the maximal number of linearly independent vectors $h_1, \cdots, h_k \in \operatorname{Ker} A$ satisfying $(Th_i, h_j) = 0$ for $i \neq j$ and $(Th_i, h_i) < 0$ for $i = 1, 2, \cdots, k$.

In fact, it is possible to construct at least one "albedo operator" E, which is a bounded strictly positive self-adjoint operator on H_T, such that

$$\psi(x) = e^{-xT^{-1}A}PE\phi_+ + (I-P)E\phi_+ \tag{2.11}$$

is a solution of (2.8)–(2.10). Here P is the continuous extension from $D(T)$ to H_T of the projection of H onto $(T[Z_0])^\perp$ along Z_0 (cf. (2.1)), while

$$\|I - E\|_{H_T} < 1 \tag{2.12}$$

(cf. [13], where it is shown that $\sigma(E) \subset (0,2)$). Evidently we must then have $(I-P)E\phi_+ \in \operatorname{Ker} A$ for all $\phi_+ \in Q_+[H_T]$.

The solution of the existence problem for (1.1)–(1.3) is provided by the following

THEOREM 2.3. *For every $\phi_+ \in Q_+[H_T]$ there exists at least one continuous function $\psi\colon [0,\infty) \to H_T$, which is continuously differentiable on $(0,\infty)$ and satisfies the equations*

$$T\psi'(x) = -A\psi(x), \qquad x \in \mathbb{R}_+, \tag{2.13}$$

$$Q_+\psi(0) = \mathscr{R}JQ_-\psi(0) + \phi_+, \tag{2.14}$$

$$\|\psi(x)\|_T = O(1) \qquad (x \to \infty). \tag{2.15}$$

Proof. Consider the operator

$$S_{\mathscr{R}} = I + \mathscr{R}J(I-E).$$

Because of the estimate

$$\|S_{\mathscr{R}} - I\|_{H_T} \leqq \|\mathscr{R}\|_{H_T}\|J\|_{H_T}\|I-E\|_{H_T} < 1, \tag{2.16}$$

the operator $S_{\mathscr{R}}$ is bounded and invertible on H_T. Consider the function

$$\psi(x) = e^{-xT^{-1}A}PES_{\mathscr{R}}^{-1}\phi_+ + (I-P)ES_{\mathscr{R}}^{-1}\phi_+, \qquad 0 \leqq x < \infty.$$

Then ψ is a continuous function from $[0,\infty)$ into H_T, which is bounded and continuously differentiable on $(0,\infty)$ and satisfies (2.13), since $(I-P)ES_{\mathscr{R}}^{-1}\phi_+ \in \operatorname{Ker} A$. Notice that $S_{\mathscr{R}}$ maps $Q_+[H_T]$ onto itself. We now have

$$\begin{aligned}(Q_+ - \mathscr{R}JQ_-)\psi(0) &= (Q_+ - \mathscr{R}JQ_-)ES_{\mathscr{R}}^{-1}\phi_+ \\ &= (Q_+ - \mathscr{R}JQ_-)EQ_+S_{\mathscr{R}}^{-1}\phi_+ = (Q_+ - \mathscr{R}JQ_-EQ_+)S_{\mathscr{R}}^{-1}\phi_+ \\ &= (Q_+ + \mathscr{R}JQ_-(I-E)Q_+)S_{\mathscr{R}}^{-1}\phi_+ = Q_+S_{\mathscr{R}}Q_+S_{\mathscr{R}}^{-1}\phi_+ = \phi_+,\end{aligned}$$

and therefore (2.14) is satisfied. □

Remark. As far as one considers only operators A which do not have negative definite parts and whose kernel is finite-dimensional, the most general boundary condition to be imposed at infinity reads

$$\exists n \geqq 0: \|\psi(x)\|_T = 0(x^n)(x \to \infty). \tag{2.17}$$

This is the case for conservative neutron transport [16] and for the Fokker–Planck equation [4], where the root subspace as defined by (2.1) is two-dimensional and $n=1$. This implies that for large x the solution behaves as $f_1 + f_2 x$. (Here the vectors f_1 and f_2 are functions depending on the angular (for neutron transport) or velocity (for the Fokker–Planck equation) variable, but in more general cases they may contain some other variables as well, depending on the complexity of the operators T and A.) For the kinetic (transport) problems usually occurring in physical situations the solution $f_1 + f_2 x$ is called normal (or Chapman–Enskog) and the vectors in the root subspace are related to the (reduced) hydrodynamical description, valid far from the boundary. Because the boundary condition (2.17) is more general than (2.15), existence of solutions is clear. For the two types of boundary condition at infinity the number of linearly independent solutions might be different if normal solutions occur.

Let us now define P as the projection onto $(T[Z_0])^{\perp}$ along Z_0 and PP_+ (resp. PP_-) as the projection onto the maximal positive (resp. negative) S-invariant subspace along the direct sum of Z_0 and the maximal negative (resp. positive) S-invariant subspace. Here positivity and negativity relate to the inner product (2.6) and essential use has been made of the Spectral Theorem for S (cf. Proposition 2.1). As a consequence,

$$(TPP_+h,h) = (SPP_+h,h)_A \geqq 0, \qquad (TPP_-k,k) = (SPP_-k,k)_A \leqq 0, \tag{2.18}$$

where strict positivity and negativity hold for $h \in \operatorname{Ran} PP_+$ and $k \in \operatorname{Ran} PP_-$. Then PP_+ and PP_- extend to bounded projections on H_T (cf. [1],[13]). Next put

$$M_{\mp,} = [\operatorname{Ran} PP_{\pm} \oplus \operatorname{Ker}(Q_{\pm} - \mathscr{R}JQ_{\mp})] \cap Z_0 \tag{2.19}$$

for the notions concerning indefinite inner product spaces we are going to use we refer to [5].

LEMMA 2.4. *We have*

$$(Th,h) \leqq 0, \qquad h \in M_{-,\mathscr{R}}.$$

If $\|\mathscr{R}\|_{H_T} < 1$, *or under the weaker assumption*

$$\operatorname{Ker}(Q_+ - \mathscr{R}JQ_-) \cap Z_0 = \{0\}, \tag{2.20}$$

we have

$$(Th,h) < 0, \qquad 0 \neq h \in M_{-,\mathscr{R}}.$$

Proof. For $h \in M_{-,\mathscr{R}}$ we first determine $g \in \operatorname{Ran} PP_+$ and $k \in \operatorname{Ker}(Q_+ - \mathscr{R}JQ_-)$ such that $h = g + k$. Since $h \in Z_0$ and $g \in (T[Z_0])^\perp$, we have $(Th, g) = 0$. Hence,

$$(Th,h)+(Tg,g)=(Tk,k)=\|Q_+k\|_T^2-\|Q_-k\|_T^2$$
$$=\|\mathscr{R}JQ_-k\|_T^2-\|Q_-k\|_T^2\leqq-\left(I-\|\mathscr{R}\|_{H_T}^2\right)\|Q_-k\|_T^2.$$

Since $(Tg,g)=(Sg,g)_A \geqq 0$ (cf. (2.18)), we have $(Th,h)\leqq 0$. Moreover, if $(Th,h)=0$, then $g=0$ and either $\|\mathscr{R}\|_{h_T}=1$ or $Q_-k=0$; the latter would imply $k=\mathscr{R}JQ_-k+Q_-k = 0$ and $h=0$. Hence, if $\|\mathscr{R}\|_{H_T}<1$, we have $(Th,h)<0$ for $0\neq h\in M_{-,\mathscr{R}}$. The latter conclusion can also be drawn under the weaker assumption (2.20). □

In the same way we can prove that

$$(Th,h)<0, \qquad h\in M_{-,\mathscr{R}}\cap \operatorname{Ker}A,$$

provided

$$\operatorname{Ker}(Q_+-\mathscr{R}JQ_-)\cap\operatorname{Ker}A=\{0\}. \tag{2.21}$$

THEOREM 2.5. *Under the condition* (2.21) *the number of linearly independent solutions of the homogeneous* $(\phi_+=0)$ *problem* (2.13)–(2.15) *coincides with the maximal number of linearly independent vectors* $h_1,\cdots,h_k\in\operatorname{Ker}A$ *satisfying* $(Th_i,h_j)=0$ *for* $i\neq j$ *and* $(Th_i,h_i)<0$ *for* $i=1,2,\cdots,k$.

Proof. Denoting by $\mathscr{R}^*$ the adjoint of $\mathscr{R}$ in H, we easily compute

$$(T[M_{\mp,\mathscr{R}}])^\perp=\left[(T\operatorname{Ran}PP_\pm)^\perp\cap\operatorname{Ran}T^{-1}\left(Q_\pm-Q_\mp J\mathscr{R}^*\right)\right]+(T[Z_0])^\perp$$
$$=\left[(\operatorname{Ran}PP_\mp\oplus Z_0)\cap\operatorname{Ran}\left(Q_\pm+Q_\mp J(\mathscr{R}^\dagger)^*\right)T^{-1}\right]+(T[Z_0])^\perp$$
$$=\left\{\left[\operatorname{Ran}PP_\mp\oplus\operatorname{Ran}\left(Q_\pm+Q_\mp J(\mathscr{R}^\dagger)^*\right)\right]\cap Z_0\right\}\cap(T[Z_0])^\perp.$$

Here we have used the intertwining property $T\mathscr{R}=\mathscr{R}^\dagger T$ and the fact that the operator $Q_\pm+Q_\mp J(\mathscr{R}^\dagger)^*$ is a bounded projection on H and therefore has closed range. For $g\in\operatorname{Ran}PP_-$ and $k=(Q_++Q_-J(\mathscr{R}^\dagger)^*)l$ we obtain

$$(Th,h)+(Tg,g)=(Tk,k)=\|Q_+k\|_T^2-\|Q_-k\|_T^2=\|Q_+l\|_T^2-\left\|Q_-J(\mathscr{R}^\dagger)^*l\right\|_T^2$$
$$\geqq\left(1-\left\|(\mathscr{R}^\dagger)^*\right\|_{H_T}^2\right)\|Q_+l\|_T^2\geqq 0,$$

because $(\mathscr{R}^\dagger)^*$ is a contraction in H_T:

$$0\leqq\left((\mathscr{R}^\dagger)^*h,h\right)_T=\left((Q_+-Q_-)h,\mathscr{R}(Q_+-Q_-)h\right)_T\leqq\|\mathscr{R}\|_{H_T}^2\|h\|_T^2\leqq\|h\|_T^2.$$

We now obtain

$$(Th,h)\geqq 0, \qquad h\in(T[M_{-,\mathscr{R}}])^\perp\cap Z_0.$$

Since Z_0 is nondegenerate with respect to the indefinite inner product

$$[h,k]=(Th,k), \tag{2.22}$$

$M_{-,\mathscr{R}}$ is negative and $(T[M_{-,\mathscr{R}}])^\perp\cap Z_0$ is positive, the subspace $M_{-,\mathscr{R}}$ is maximal negative with respect to this inner product. Under the condition (2.21) the subspace $M_{-,\mathscr{R}}\cap\operatorname{Ker}A$ then is strictly negative and maximal in this respect among the subspaces

of $\mathrm{Ker}\,A$. Because the linear span of the vectors $h_1,\cdots,h_k$ in the statement of this theorem is also a maximal strictly negative subspace of $\mathrm{Ker}\,A$ and the dimension of such a subspace does not depend on its specific choice, we must have $\dim(M_{-,\mathscr{R}}\cap \mathrm{Ker}\,A)=k$. Finally, if ψ is a solution of (2.13)–(2.15) with $\phi_+=0$, then necessarily $(I-P)\psi(0)\in M_{-,\mathscr{R}}\cap \mathrm{Ker}\,A$. □

Under the condition (2.21) we find the same existence and uniqueness result as for $\mathscr{R}=0$, which we easily see on comparing Theorems 2.2 and 2.5. If one would drop condition (2.21), the homogeneous ($\phi_+=0$) problem (2.13)–(2.15) in general has more linearly independent solutions than is to be expected from the above theorem. As an example, consider the case $\mathscr{R}=I$, which describes *purely specular reflection*. First we observe that every $k\in H_T$, which satisfies $Q_+k=\mathscr{R}JQ_-k$ for $\mathscr{R}=I$, has the property

$$(Tk,k)=\|Q_+k\|_T^2-\|Q_-k\|_T^2=\|JQ_-k\|_T^2-\|Q_-k\|_T^2=0,$$

since J is a unitary operator on H_T. If such a vector k would belong to the space $\mathrm{Ran}\,PP_+\oplus\mathrm{Ker}\,A$, as it should be if it were the initial value of a solution ψ, then $k=-g+h$ for some $g\in\mathrm{Ran}\,PP_+$ and $h\in\mathrm{Ker}\,A$. Since $Q_+k=JQ_-k$ implies

$$Jk=JQ_+k+JQ_-k=J(JQ_-k)+Q_+k=k,$$

and therefore $Jg=g$ and $Jh=h$, the property $g=Jg\in\mathrm{Ran}\,PP_-$ would give rise to $g=0$ and thus $k\in\mathrm{Ker}\,A$, whence $k\in\mathrm{Ker}\,A\cap\mathrm{Ker}(I-J)$. Conversely, every such k would fulfill the condition $JQ_-k=Q_+Jk=Q_+k$ and therefore be an initial value of some solution ψ. Thus the constant functions $\psi(x)=k$, where $k=Jk\in\mathrm{Ker}\,A$, are the solutions of the homogeneous ($\phi_+=0$) problem (2.13)–(2.15) with $\mathscr{R}=I$.

Remark. The analogue of Theorem 2.5 for the kinetic equation (2.13) with boundary conditions (2.14) and (2.17) can easily be obtained by repeating the arguments with Z_0 instead of $\mathrm{Ker}\,A$. It then appears that under the assumption (2.20) the number of linearly independent solutions of the homogeneous ($\phi_+=0$) problem coincides with the maximal number of linearly independent vectors $h_1,\cdots,h_k\in Z_0$ satisfying $(Th_i,h_j)=0$ for $i\neq j$ and $(Th_i,h_i)<0$ for $i=1,2,\cdots,k$, which is the same result as for $\mathscr{R}=0$. In the case of purely specular reflection ($\mathscr{R}=I$) this number generally is larger and in fact equals the dimension of the subspace $Z_0\cap\mathrm{Ker}(I-J)$ of "even" root subspace vectors. In general, for $Z_0\neq\mathrm{Ker}\,A$ one will find a larger measure of nonuniqueness of the solution than for the problem (2.13)–(2.15), which can be accounted for by considering the normal solutions f_1+f_2x.

3. An iteration procedure. Let us consider a suitable bounded strictly positive albedo operator E on H_T, such that $\psi(0)=E\phi_+$ yields a solution of (2.8)–(2.10). Such an operator always exists and satisfies (2.12). It is unique, if and only if $(Th,h)\geqq 0$ for all $h\in\mathrm{Ker}\,A$ (cf. Theorem 2.2). Using the norm estimate (2.16), we may write a solution of (2.13)–(2.15) as follows:

$$\psi(x)=e^{-xT^{-1}A}PEg_+ +(I-P)Eg_+,\qquad x\in\mathbb{R}_+,$$

where

$$g_+=S_{\mathscr{R}}^{-1}\phi_+=\sum_{n=0}^{\infty}(-1)^n[\mathscr{R}j(I-E)]^n\phi_+; \tag{3.1}$$

the series is absolutely convergent in the norm of H_T, uniformly in ϕ_+ on bounded subsets of $Q_+[H_T]$. We may therefore compute g_+ by iterating the vector equation

$$g_+ +\mathscr{R}J(I-E)g_+=\phi_+ \tag{3.2}$$

on $Q_+[H_T]$. Depending on the choice of the albedo operator E—unique if and only if $(Th,h)\geqq 0$ for $h\in \operatorname{Ker}A$—, different solutions are generated. In order to find all solutions, especially in the cases where they are nonunique, one should still solve the homogeneous ($\phi_+=0$) problem (2.13)–(2.15). For instance, if $\operatorname{Ker}A\neq\{0\}$ and $Jh=h$ for all $h\in\operatorname{Ker}A$, which occurs for the Fokker–Planck example (1.5)–(1.7) (disregarding for the moment that this model does not satisfy the boundedness assumption on A), we would have $(Th,h)=(TJh,Jh)=-(JTh,Jh)=-(Th,h)=0$ for all $h\in\operatorname{Ker}A$. This would imply existence of a unique albedo operator and therefore the generation by iteration of one solution only. Nevertheless the problem is nonuniquely solvable and the homogeneous problem should be solved as well. A similar remark applies to the solution of (2.13) with boundary conditions (2.14) and (2.17).

Let us consider the case when A is a compact perturbation of the identity satisfying

$$\exists 0<\alpha<1:\operatorname{Ran}(I-A)\subset\operatorname{Ran}|T|^{\alpha},\qquad Z_0\subset D\big(|T|^{2+\alpha}\big), \tag{3.3}$$

which occurs in one-speed and symmetric multigroup neutron transport (cf. [19]), and several BGK models in rarefied gas dynamics. If we choose a closed subspace $\mathbb{B}\supset\operatorname{Ran}(I-A)$, which may be chosen finite-dimensional if $I-A$ has finite rank, and operators $\pi:H\to\mathbb{B}$ and $j:\mathbb{B}\to H$ such that πj is the identity on $\mathbb{B}$ and $j\pi$ the orthogonal projection of H onto $\mathbb{B}$, a representation for E can be found in terms of generalized Chandrasekhar **H**-functions. More precisely, if $\sigma(\cdot)$ denotes the resolution of the identity of the self-adjoint operator T, we have (see [21])

$$E\phi_+=\phi_+ +\int_{-\infty}^{0}\int_{0}^{\infty}\frac{\nu}{\nu-\mu}\sigma(d\mu)(I-A)j\mathbf{H}_l(-\mu)\mathbf{H}_r(\nu)\pi\sigma(d\nu)\phi_+, \tag{3.4}$$

where $\mathbf{H}_l(-\mu)$ and $\mathbf{H}_r(\nu)$ are solutions of the nonlinear integral equations

$$\mathbf{H}_l(z)^{-1}=I-z\int_0^{\infty}(z+t)^{-1}\mathbf{H}_r(t)\pi\sigma(dt)(I-A)j, \tag{3.5}$$

$$\mathbf{H}_r(z)^{-1}=I-z\int_0^{\infty}(z+t)^{-1}\pi\sigma(-dt)(I-A)j\mathbf{H}_l(t). \tag{3.6}$$

The solutions and their inverses must be analytic for $\operatorname{Re}z>0$ and continuous for $\operatorname{Re}z\geqq 0$. If $\operatorname{Ker}A\neq\{0\}$, the continuity of $\mathbf{H}_l$ and $\mathbf{H}_r$ at infinity must be replaced by a weaker requirement. (The precise description of such requirements was not given in [21].) Equation (3.2) then has the form

$$g_+-\int_{\substack{-\infty\\(\mu)}}^{0}\int_{\substack{0\\(\nu)}}^{\infty}\frac{\nu}{\nu-\mu}\mathscr{R}J\sigma(d\mu)(I-A)j\mathbf{H}_l(-\mu)\mathbf{H}_r(\nu)\pi\sigma(d\nu)g_+=\phi_+. \tag{3.7}$$

On solving the **H**-equations (3.5)–(3.6) we may compute g_+ by iteration. It should be noted that the above expression (3.4) for E was formulated for $\phi_+\in Q_+[H]$, but allows continuous extension to $\phi_+\in Q_+[H_T]$.

Let us consider the specific example of the scalar BGK model. The existence and uniqueness theory for this example without reflection is immediate from [1], and has also been published by Kaper [15]. For a combination of specular and diffuse reflection (no absorption) solutions were obtained before by Cercignani [8], using expansion with respect to increasing powers of the accommodation coefficient α. Let $L_2(\mathbb{R})_\delta$ be the Hilbert space of complex measurable functions on $\mathbb{R}$ with inner product

$$(h,k)=\int_{-\infty}^{\infty}h(v)\overline{k(v)}\,d\delta(v),\qquad d\delta(v)=\pi^{-1/2}e^{-v^2}dv,$$

and define T, Q_+, Q_-, A, $\mathscr{R}$ and J as follows:

$$(Th)(v)=vh(v), \qquad (Ah)(v)=h(v)-\pi^{-1/2}\int_{-\infty}^{\infty} h(v')e^{-(v')^2}dv',$$

$$(Q_+h)(v)=\begin{cases} h(v), & v>0,\\ 0, & v<0,\end{cases} \qquad (Q_-h)(v)=\begin{cases} 0, & v>0,\\ h(v), & v<0,\end{cases}$$

$$(Jh)(v)=h(-v), \qquad (\mathscr{R}h)(v)=\alpha h(v)+2\beta\pi^{-1/2}\int_0^{\infty} v'h(v')e^{-(v')^2}dv',$$

where $\mathscr{R}$ is defined on $\operatorname{Ran} Q_+$ and $\alpha,\beta\geqq 0$ with $\alpha+\beta\leqq 1$. This model satisfies the assumptions of the previous section and existence is assured. First we solve the **H**-equation (i.e., (3.5)–(3.6) with $\mathbf{H}_l=\mathbf{H}_r$ and $\mathbb{B}=\{\text{constant functions}\}$)

$$\mathbf{H}(z)^{-1}=1-\frac{z}{\sqrt{\pi}}\int_0^{\infty}(z+t)^{-1}\mathbf{H}(t)e^{-t^2}dt,$$

requiring a solution such that $\mathbf{H}$ and $\mathbf{H}^{-1}$ are analytic for $\operatorname{Re} z>0$, continuous for $\operatorname{Re} z\geqq 0$ and satisfying $\mathbf{H}(z)=O(z)$ for $z\to\infty$ with $\operatorname{Re} z\geqq 0$. We find

$$(E\phi_+)(v)=\begin{cases}\phi_+(v), & v>0,\\ \dfrac{1}{\sqrt{\pi}}\displaystyle\int_0^{\infty}\frac{v'}{b'-v}\mathbf{H}(-v)\mathbf{H}(v')\phi_+(v')e^{-(v')^2}dv', & v<0.\end{cases}$$

Therefore, we write (3.2) in the form

$$g_+(v)-\frac{\alpha}{\sqrt{\pi}}\int_0^{\infty}\frac{v'}{v'+v}\mathbf{H}(v)\mathbf{H}(v')g_+(v')e^{-(v')^2}dv'$$

$$-\frac{2\beta}{\pi}\int_0^{\infty}\int_0^{\infty}\frac{vv'}{v'+v}\mathbf{H}(v)\mathbf{H}(v')g_+(v')e^{-[v^2+(v')^2]}dv'\,dv=\phi_+(v),$$

which has to be solved by iteration. The initial value of the solution is then given by $\psi(0,v)=g_+(v)$ for $v>0$ and by

$$\psi(0,v)=\frac{1}{\sqrt{\pi}}\int_0^{\infty}\frac{v'}{v'-v}\mathbf{H}(-v)\mathbf{H}(v')g_+(v')e^{-(v')^2}dv'$$

for $v<0$.

For the isotropic Lorentz gas (neutron transport) the calculation has been carried out for various combinations of selective, specular and diffuse boundary conditions (cf. [10], [11]) yielding interesting and sometimes striking conclusions about their influence on the boundary layer structure, density profile at the wall, validity of Fick's law, etc. For instance, the selective reflection of slow particles and absorption of fast ones leads to an accumulation of particles near the wall and a reversal of the density gradient (interpreted in terms of Fick's law, as equivalent to a negative diffusion coefficient [11]). For the Fokker–Planck equation such selective boundary conditions have been investigated numerically by Burschka and Titulaer [6]. We remark that in general (e.g. for so-called selective boundary conditions [6], [11]) the operator $\mathscr{R}$ is not self-adjoint in H_T. Since only the contraction property of $\mathscr{R}$ plays a role in all derivations and not whether it is self-adjoint, our existence, uniqueness and iteration results also apply to selective boundary conditions.

4. Discussion.

4.1. Generalization to unbounded A. Hitherto we have assumed that A is a bounded operator. For many applications, especially the ones involving BGK models in rarefied gas dynamics, this is sufficient. The Fokker–Planck model (1.5)–(1.7), however, does not satisfy these assumptions. We shall therefore point out what type of hypotheses on T and A, with A unbounded, would entail a repetition of the previous arguments.

Let us assume that T is (bounded or unbounded) injective self-adjoint on H, and let us define Q_+, Q_- and H_T as before. Suppose A is a positive self-adjoint operator with closed range and finite-dimensional kernel, possibly unbounded, such that $D(T) \cap D(A)$ is dense in H. On defining Z_0 as before and repeating the previous hypotheses on Z_0, we may derive Proposition 2.1. It should be noted that $T^{-1}A$, with $D(T^{-1}A) = \{h \in D(A)/Ah \in \operatorname{Ran} T\}$, is closable, but not necessarily closed. Still we may derive the decompositions (2.2) and (2.3) where $T^{-1}A|_{Z_0}$ is closed (and even bounded) and $T^{-1}A$ and S^{-1} should be replaced by their respective closures. As a result, S will be a closed *symmetric* operator with respect to the inner product (2.6). We shall assume that S is, in fact, self-adjoint on the completion of $(T[Z_0])^{\perp} \cap D(A)$ (which is dense in H, due to the density of $D(T) \cap D(A)$) with respect to (2.6). By H_A we shall denote the direct sum of this completion and Z_0. Since A has a closed range and a finite-dimensional kernel, H_A is densely imbedded in H. We define H_K as the direct sum of Z_0 and the completion of $D(T) \cap H_A (\supset D(T) \cap D(A))$ with respect to the inner product

$$(h,k)_K = (|S|h,k)_A,$$

where the absolute value of S is taken in H_A. As before, we define the projections P, PP_+ and PP_- on H_A (and not on H) and extend them continuously to projections on H_K (and not on H_T).

If A is bounded, we may identify H_A and H (which is a trivial observation) as well as H_K and H_T (see [1]; cf. [13] for a different proof). The existence and uniqueness theory has then been developed in §2. For a large class of models on $L_2(a,b)$, where T is a multiplication by an indefinite weight function and A is a Sturm–Liouville type differential operator, it has been proved by Beals [2] that the Hilbert spaces H_T and H_K are completions of $D(T) \cap D(A)$ with respect to equivalent inner products and can be identified. Moreover, for the models Beals considered the previous assumptions on T and A, including the self-adjointness assumption on S, are satisfied. As for bounded A, we may then develop the theory of §2 and the first paragraph of §3 for these indefinite Sturm–Liouville problems and essentially the same results are found. Moreover, for these cases the operator S is bounded self-adjoint on $H_A \cap (T[Z_0])^{\perp}$ (which is due to more specific assumptions on T and A) and even compact. A specific example of such a model is the Fokker–Planck equation (1.5)–(1.7). For this example the equivalence proof of H_T and H_K is contained in [3]. It should be noticed that Theorem 2.3 answers in the affirmative the existence issue raised in [4], thereby making redundant the condition imposed there to enforce existence of solutions (namely, the condition $Bl \in \operatorname{cls}\{Bu_n : n > 0\}$ in [4]).

4.2. The albedo operator for indefinite Sturm–Liouville problems. It is by no means clear how to proceed finding the albedo operator E for (1.5)–(1.7) and other Sturm–Liouville type models. One way, suggested by the approach in [4], is to use the completeness of the eigenfunctions $(u_n)_{0 \neq n \in \mathbb{Z}}$ of $T^{-1}A$ at the nonzero eigenvalues $(\lambda_n)_{0 \neq n \in \mathbb{Z}}$, where we order these by $\ldots \leqq \lambda_{-2} \leqq \lambda_{-1} < 0 < \lambda_1 \leqq \lambda_2 \leqq \ldots$ and take

account of multiplicities. (It should be noted that under weak oscillation conditions on A these eigenvalues are simple). We add an orthogonal basis $u_{0,1}, \cdots, u_{0,l}$ of a given maximal positive subspace N_+ of $\operatorname{Ker} A$ (i.e., $(Tu_{0,i}, u_{0,i}) \geqq 0$). The full-range completeness property implies that every vector $h \in PP_+[H_K] \oplus N_+$ can be expanded as the series

$$h = \sum_{i=1}^{l} \xi_{0,i} u_{0,i} + \sum_{n=1}^{\infty} \xi_n u_n. \tag{4.1}$$

Half-range completeness (for the problem without reflection) amounts to the possibility of expanding every vector $g_+ \in Q_+[H_T]$ (where $H_K \simeq H_T$) as

$$g_+ = \sum_{i=1}^{l} \eta_{0,i} Q_+ u_{0,i} + \sum_{n=1}^{\infty} \eta_n Q_+ u_n;$$

here Q_+ is the restriction to the interval I_+ where the indefinite weight is positive. (For (1.5)–(1.7) we have $I_+ = \mathbb{R}_+$). *Assuming* the existence of a nonnegative weight function $\mathbf{H}$ on I_+ satisfying

$$\int_{I_+} u_{0,i}(v) u_{0,j}(v) \mathbf{H}(v)\, dv = \delta_{i,j} \theta_{0,i} \qquad \theta_{0,i} > 0, \tag{4.2}$$

$$\int_{I_+} u_n(v) u_m(v) \mathbf{H}(v)\, dv = \delta_{n,m} \theta_n, \qquad \theta_n > 0, \tag{4.3}$$

$$\int_{I_+} u_{0,i}(v) u_n(v) \mathbf{H}(v)\, dv = 0, \tag{4.4}$$

we can easily evaluate the (unique) albedo operator E such that

$$EQ_+[H_T] = PP_+[H_K] \oplus N_+.$$

Indeed, on expanding $h = Eg_+$ with $g_+ \in Q_+[H_T]$ as the series (4.1) we obtain

$$g_+ = Q_+ E g_+ = \sum_{i=1}^{l} \xi_{0,i} Q_+ u_{0,i} + \sum_{n=1}^{\infty} \xi_n Q_+ u_n. \tag{4.5}$$

Using (4.2)–(4.4), we then easily derive

$$Eg_+ = \sum_{i=1}^{l} \theta_{0,i}^{-1} \left(\int_{I_+} u_{0,i}(v) g_+(v) \mathbf{H}(v)\, dv \right) u_{0,i} + \sum_{n=1}^{\infty} \theta_n^{-1} \left(\int_{I_+} u_n(v) g_+(v) \mathbf{H}(v)\, dv \right) u_n.$$

If the weight function $\mathbf{H}$ on I_+ is known, the boundary value problem with reflection can again be solved by iterating (3.2), using (4.6). At present even the existence (let alone the computation) of such a weight function is an open problem.

REFERENCES

[1] R. BEALS, *On an abstract treatment of some forward-backward problems of transport and scattering*, J. Funct. Anal., 34 (1979), pp. 1–20.

[2] ———, *Indefinite Sturm–Liouville problems and half-range completeness*, J. Differential Equations, to appear.

[3] R. BEALS AND V. PROTOPOPESCU, *Half-range completeness for the Fokker–Planck equation*, J. Stat. Phys., 32 (1983), pp. 565–584.

[4] _____, *Half-range completeness and orthogonality for the Fokker–Planck equation with general boundary conditions*, II, Transp. Theor. Stat. Phys., 13 (1984), pp. 43–55.

[5] J. BOGNÁR, *Indefinite Inner Product Spaces*, Springer, Berlin, 1974.

[6] M. A. BURSCHKA AND U. M. TITULAER, *The kinetic boundary layer for the Fokker–Planck equation: selectively absorbing boundaries*, J. Stat. Phys., 26 (1981), pp. 59–71.

[7] K. M. CASE AND P. F. ZWEIFEL, *Linear Transport Theory*, Addison-Wesley, Reading, MA, 1967.

[8] C. CERCIGNANI, *The Kramers problem for a not completely diffusing wall*, J. Math. Anal., 10 (1965), pp. 568–586.

[9] S. CHANDRASEKHAR, *Radiative Transfer*, Oxford Univ. Press, London, 1950; Dover, New York, 1960.

[10] R. G. COLE, T. KEYES AND V. PROTOPOPESCU, *Stationary transport with partially reflecting boundary conditions*, J. Chem. Phys., 81 (1984), pp. 2771–2775.

[11] _____, *Stationary transport with partially reflecting boundary conditions* II, in preparation.

[12] W. GREENBERG AND C. V. M. VAN DER MEE, *An abstract model for radiative transfer in an atmosphere with reflection by the planetary surface*, this Journal, in press.

[13] W. GREENBERG, C. V. M. VAN DER MEE AND P. F. ZWEIFEL, *Generalized kinetic equations*, Integral Equations Operator Theory, 7 (1984), pp. 60–95.

[14] R. J. HANGELBROEK, *Linear analysis and solution of neutron transport problems*, Transport Theory Statist. Phys., 5 (1976), pp. 1–85.

[15] H. G. KAPER, *A constructive approach to the solution of a class of boundary value problems of mixed type*, J. Math. Anal. Appl., 63 (1978), pp. 691–718.

[16] C. G. LEKKERKERKER, *The linear transport equation. The degenerate case $c=1$*, II. *Half-range theory*, Proc. Royal Soc. Edinburgh, 75A(1976), pp. 283–295.

[17] N. B. MASLOVA, *The Kramers problem in the kinetic theory of gases*, USSR Comput. Math. and Math. Phys., 22 (1982), pp. 208–219.

[18] _____, *Stationary solutions of the linearized Boltzmann equation*, Trudy Matem. Instituta im. V. A. Steklova, 159 (1983), pp. 41–60. (In Russian).

[19] C. V. M. VAN DER MEE, *Semigroup and factorization methods in transport theory*, Math. Centre Tract No. 146, Math. Centre, Amsterdam, 1981.

[20] _____, *Abstract boundary value problems modeling transport processes in semi-infinite geometry*, Transp. Theor. Stat. Phys., 13 (1984), pp. 29–41.

[21] _____, *Albedo operators and H-equations for generalized kinetic models*, Transp. Theor. Stat. Phys., 13 (1984), pp. 341–376.

SIAM J. MATH. ANAL.
Vol. 17, No. 4, July 1986

013

ON A PROBLEM IN THE POLYMER INDUSTRY: THEORETICAL AND NUMERICAL INVESTIGATION OF SWELLING*

A. FASANO†, G. H. MEYER‡ AND M. PRIMICERIO†

Abstract. A recent model for the penetration of solvents into polymers leads to a parabolic free boundary problem with unusual boundary conditions. It is shown that the model equations are well posed, and some qualitative features of the free boundary are established. A numerical method for the free boundary problem is suggested and its convergence is proved. A numerical calculation is included to illustrate the theoretical results.

1. Introduction. This paper deals with a mathematical model proposed in [2] for the penetration of solvents into polymers. Further comments on this model may be found in [1], [3] and [4].[1] A related but more involved and mathematically unresolved model for swelling may be found in [10] where multiple phases are allowed.

While we confine our analysis to the particular case of constant solvent concentration at the polymer surface, the one-dimensional theory presented here is comprehensive and includes existence, uniqueness, regularity and other qualitative properties of the solution, asymptotic estimates and a convergent numerical algorithm. These analytic results may prove useful for model verification.

Let us sketch the physical problem.

Consider a slab of a glassy polymer (e.g., poly (methyl methacrylate)) in contact with a solvent (n-alkyl alcohol). It is observed that if the solvent concentration exceeds some threshold value, then the solvent moves into the polymer, creating a swollen layer in which the solvent diffuses according to Fick's law. The boundary between the swollen region and the glassy region obeys an empirical penetration law, relating its velocity with the (unknown) value assumed on it by the solvent concentration. An additional condition on the free boundary is obtained imposing mass conservation, i.e., equating the mass density current to the product of the solvent concentration and the velocity of the free boundary.

Assuming that the swelling process occurs instantaneously at the penetration surface and choosing a frame of reference in which the swollen region is at rest, the above scheme leads to the following statement.

Problem (P). *Find a triple* (T,s,c) *such that* $T>0$, $s\in C^1[0,T]$, $c\in C^{2,1}(D_T)\cap C(\bar{D}_T)$, $D_T=\{(x,t): 0<x<s(t),\ 0<t<T\}$, c_x *continuous up to* $x=s(t)$, *and such that*

$$c_{xx}-c_t=0, \qquad (x,t)\in D_T, \tag{1.1}$$

$$s(0)=0, \tag{1.2}$$

$$c(0,t)=1, \qquad 0<t<T, \tag{1.3}$$

$$\dot{s}(t)=f[c(s(t),t)], \qquad 0<t<T, \tag{1.4}$$

$$c_x(s(t),t)=-\dot{s}(t)[c(s(t),t)+q], \qquad 0<t<T. \tag{1.5}$$

*Received by the editors July 11, 1984, and in revised form March 28, 1985. This work was partly supported by USA-ERO under contract DAJA 45-83-C-0053 and the National Science Foundation under grant MCS 8302548.

†Universita Degli Studi, Istituto Matematico, Ulisse Dini, Viale Morgagni, 57/A, I 50134 Firenze, Italy

‡School of Mathematics, Georgia Institute of Technology, Atlanta, Georgia 30332.

[1]Some preliminary results were communicated at the conference on Applied Nonlinear Functional Analysis, Berlin 1981 and at the Seminar on Mathematics and Industry, Oberwolfach 1983.

In (1.1)–(1.5) nondimensional variables are used and physical constants are normalized to 1. In (1.5) q is a given nonnegative constant, representing the threshold concentration for penetration. By $c(x,t)$ we denote the excess concentration, normalized to 1 at $x=0$, so that the normalized solvent concentration is represented by $c+q$. Condition (1.4) describes the penetration law. Throughout the paper the function f will be supposed to satisfy the following assumptions:

$$\text{(F)} \qquad f\in C^1(0,1], \qquad f'(c)>0 \quad \text{for } c\in(0,1], \qquad f(0)=0.$$

(Indeed the form for $f(c)$ proposed in [2] is αc^m, $m>0$.)

As a consequence, (1.4) can be written as

$$(1.4') \qquad c(s(t),t)=\Phi(\dot{s}(t)),$$

where $\Phi=f^{-1}$ also satisfies (F).

The plan of the paper is the following. In §2 a priori estimates are obtained for a class of auxiliary problems. Such estimates are used in §3 to prove local existence of a smooth solution. A general uniqueness theorem is then proved in the same section. In §4 it is proved that the solution can be continued over an arbitrarily large time interval, that it depends continuously on the data f,q, and that the free boundary is a C^2 convex curve. The asymptotic behavior of the free boundary is investigated in §5, showing its crucial dependence on q and on the modulus of continuity of $f(c)$ near $c=0$. Section 6 is devoted to setting up a numerical scheme based on the sweep method introduced in [8] for a time discretization of (1.1)–(1.5). Numerical results are presented in §7.

Of course, it makes sense to consider different boundary conditions for the model studied here and also to investigate different models which have been proposed for the same phenomenon. This will be the objective of further research.

2. An auxiliary problem. Let $r(t)\in C^1[0,T]\cap C^2(0,T)$ be such that

$$(2.1) \qquad r(0)=0,$$

$$(2.2) \qquad \dot{r}(0)=f(1),$$

$$(2.3) \qquad 0\leqq\dot{r}(t)\leqq f(1) \quad \text{in } [0,T],$$

$$(2.4) \qquad |\ddot{r}(t)|\leqq K \quad \text{in } (0,T),$$

and consider the problem of finding $c\in C^{2,1}(D)\cap C(\bar{D})$, c_x continuous up to $x=r(t)$, $t\in(0,T)$, such that

$$(2.5) \qquad c_{xx}-c_t=0 \quad \text{in } D\equiv\{(x,t): 0<x<r(t), 0<t<T\},$$

$$(2.6) \qquad c(0,t)=1, \qquad 0<t<T,$$

$$(2.7) \qquad c_x(r(t),t)=-\dot{r}(t)[q+\Phi(\dot{r}(t))], \qquad 0<t<T.$$

We have immediately

PROPOSITION 2.1. *Problem* (2.5)–(2.7) *has a unique solution.*

Proof. See, e.g. [5]. □

Next, we prove

PROPOSITION 2.2. *Under assumptions* (2.1)–(2.4), *the solution of* (2.5)–(2.7) *is such that*

$$(2.8) \qquad 1>c(x,t), \qquad 0<x\leqq r(t), \quad 0<t\leqq T,$$

$$(2.9) \qquad 0>c_x(x,t)>-f(1)[q+1] \quad \text{in } \bar{D}\setminus\{0,0\}.$$

Proof. To prove (2.8), it suffices to recall the maximum principle and the sign of $c_x(r(t),t)$. Since $c(x,t)$ attains its maximum on $x=0$, we have $c_x(0,t)<0$ and the first inequality (2.8) follows from the maximum principle. Moreover, $c_{xx}(0,t)=0$. Thus c_x attains its minimal value on $x=r(t)$, and this proves (2.9). □

At this point we note $\dot{r}(t)\geqq f(1)-KT$, we fix $\dot{r}_0<f(1)$ and we reduce T, if necessary, in order to have

$$\dot{r}(t)\geqq \dot{r}_0, \qquad 0\leqq t\leqq T. \tag{2.10}$$

We prove

PROPOSITION 2.3. *Under assumptions* (2.1)–(2.4) $c\in C^{2,1}(\bar{D})$, $c_{xt}\in C(\bar{D}\setminus(0,0))$, *and there exists a constant M depending on f, and on K such that*

$$|c_t(x,t)|\leqq Mt \quad \text{in } D. \tag{2.11}$$

Proof. Note that c_t coincides with the solution $w(x,t)$ of the heat equation in D with boundary conditions

$$w(0,t)=0, \qquad 0<t<T, \tag{2.12}$$

$$[w_x+\dot{r}(t)w]_{x=r(t)}=-\ddot{r}(t)[q+\Phi(\dot{r}(t))+\Phi'(\dot{r}(t))\dot{r}(t)]\equiv F(t), \qquad 0<t<T. \tag{2.13}$$

We have that $w\in C(\bar{D})$ and that

$$\min\Big[0, \inf_{0<t<T} F(t)/\dot{r}_0\Big]\leqq w(x,t)\leqq \max\Big[0, \sup_{0<t<T} F(t)/\dot{r}_0\Big] \quad \text{in } D. \tag{2.14}$$

But this implies

$$|w_x(r(t),t)|\leqq \sup|F(t)|+f(1)\sup|F(t)|/\dot{r}_0\equiv\Psi \tag{2.15}$$

and finally

$$|w(x,t)|\leqq \Psi r(t)\leqq \Psi f(1)t \quad \text{in } D. \tag{2.16}$$

This proves (2.11). □

Now, we investigate how the solution of (2.5)–(2.7) depends on $r(t)$. Let $r_1(t)$, $r_2(t)$ satisfy (2.1)–(2.4) and let $c_1(x,t)$, $c_2(x,t)$ be the corresponding solutions of (2.5)–(2.7). We have

PROPOSITION 2.4. *Under the assumptions above, constants* $T_0>0$ *and* $N>0$ *can be found such that for any* $T\in(0,T_0)$

$$|c_1(r_1(t),t)-c_2(r_2(t),t)|\leqq N\|r_1-r_2\|_{C^1(0,T)}T, \qquad 0<t<T. \tag{2.17}$$

Proof. Let

$$\lambda(t)=\inf(r_1(t),r_2(t)),$$
$$\mu(t)=\sup(r_1(t),r_2(t))$$

and

$$D^*\equiv\{(x,t): 0<x<\lambda(t), 0<t<T\}.$$

Moreover we reduce, if necessary, T in order to have

$$\dot{r}_i(t)\geqq \dot{r}_0, \qquad t\in[0,T], \quad i=1,2.$$

The absolute value of the difference $v(x,t)=c_1(x,t)-c_2(tx,t)$ is dominated in D^* by the solution $V(x,t)$ of the heat equation with boundary data

$$V(0,t)=0, \qquad 0<t<T,$$
$$V_x(\lambda(t),t)=C\|r_1-r_2\|_{C^1(0,T)}+Mt\|r_1-r_2\|_{C[0,T]}, \qquad 0<t<T,$$

where $C=q+\max_{y\in[i_0,f(1)]}\Phi'(y)$. This implies

$$|v(\lambda(t),t)|\leqq(C+Mt)\|r_1-r_2\|_{C^1(0,T)}\cdot f(1)t, \qquad 0<t<T. \tag{2.18}$$

As a consequence of (2.18) we have in $(0,T)$

$$|c_1(r_1(t),t)-c_2(r_2(t),t)|\leqq(C+Mt)f(1)\|r_1-r_2\|_{C^1(0,T)}+f(1)(1+q)\|r_1-r_2\|_{C(0,T)}$$

thus proving (2.17). □

3. Local existence. Uniqueness. Let $\gamma(t)$ be a positive nonincreasing function defined for $t>0$ and possibly diverging for $t\to 0^+$. Denote by $X(K,T,\gamma)$ the set of functions $r(t)$ satisfying (2.1)–(2.4) and such that, for some $\alpha\in(0,1]$

$$|\dot r(t_1)-\dot r(t_2)|\leqq\gamma(\tau)(t_1-t_2)^{\alpha/2}, \qquad 0<\tau\leqq t_2\leqq t_1\leqq T. \tag{3.1}$$

Note that the set X is closed in $C^1(0,T)$.

For any $r\in X$ let c be the solution of (2.5)–(2.7). Then, define the transformation $\tilde r=\mathscr{C}r$ as follows

$$\tilde r(0)=0, \qquad \dot{\tilde r}(t)=f(c(r(t),t)), \quad 0<t<T. \tag{3.2}$$

Note that the domain of f is $[0,1]$. Thus, (3.2) makes sense only when $c(r(t),t)\geqq 0$ (remember $c<1$ by (2.8)). But from (2.9) and (2.3) we have

$$c(r(t),t)\geqq 1-[f(1)]^2(q+1)t.$$

This means that we can reduce T, if necessary, in order to have

$$c(r(t),t)\geqq c_0, \qquad 0<t<T \tag{3.3}$$

for some $c_0>0$.

Now we prove

THEOREM 3.1. *There exist K,T,γ such that the transformation $\tilde r=\mathscr{C}r$ is a contractive mapping of $X(K,T,\gamma)\subset C^1(0,T)$ into itself.*

Proof. Since (2.1)–(2.3) are satisfied by construction, to prove that $\mathscr{C}$ maps X into itself, we only need to prove that $\tilde r$ satisfies (2.4) and (3.1) for suitable K,T,γ.

We have

$$\ddot{\tilde r}(t)=[c_x(r(t),t)\dot r(t)+c_t(r(t),t)]f'(c(r(t),t)), \qquad 0<t<T. \tag{3.4}$$

Using (2.9) and (2.11), we find

$$|\ddot{\tilde r}(t)|\leqq\{[f(1)]^2(q+1)+Mt\}C_1, \qquad 0<t<T, \tag{3.5}$$

where $C_1=\max_{[c_0,1]}f'(c)$. Take e.g. $K=2C_1[f(1)]^2(q+1)$ and T such that

$$MT\leqq[f(1)]^2(q+1).$$

Then (3.4) yields

$$|\ddot{\tilde r}(t)|\leqq K, \qquad 0<t<T.$$

To estimate the Hölder norm of $\ddot{r}(t)$ we need to estimate the norm of $c_x(x,t)$ in the space $C^{1+\alpha}$. This is accomplished as follows. Define $z(x,t)=c_x(x,t)+\dot{r}(t)[q+\Phi(\dot{r}(t))]$, which solves

$$
\begin{aligned}
&z_{xx}-z_t=-\ddot{r}(t)\left[q+\Phi(\dot{r}(t))+\Phi'(\dot{r}(t))\dot{r}(t)\right] \quad \text{in } D,\\
&z_x(0,t)=0, \qquad 0<t<T,\\
&z(r(t),t)=0, \qquad 0<t<T.
\end{aligned} \tag{3.6}
$$

For any $\tau\in(0,T)$ transform the domain $0<x<r(t)$, $\tau/2<t<T$ into the rectangle $(0,1)\times(\tau/2,T)$ by the transformation $y=x/r(t)$ and apply the standard Schauder estimates (e.g. [7, Thm. 5.2, p. 561]) to the transformed function $\hat{z}(y,t)$, in the rectangle $(1/2,1)\times(\tau/2,T)$. We find

$$
\|z\|_{C^\alpha}\leqq\bar{\gamma}(\tau), \quad \text{in } \left(\tfrac{1}{2},1\right)\times(\tau,T), \tag{3.7}
$$

where $\bar{\gamma}$ depends on K, on f, on T, and on τ (and α). Thus, defining $\gamma(t)$ as suggested by (3.7), $\tilde{r}$ will satisfy (3.1). The final step in proving Theorem 3.1 is to prove the contractive character of $\mathscr{C}$. But this is an immediate consequence of (2.17) and of (3.2) because

$$
\|\tilde{r}_1-\tilde{r}_2\|_{C^1(0,T)}\leqq C_1NT\|r_1-r_2\|_{C^1(0,T)}
$$

and it suffices to reduce T, if necessary, to conclude the proof. □

Hence a $T_0>0$ can be found such that the following theorem holds.

THEOREM 3.2. *Problem* (P) *admits a solution for* $T\leqq T_0$. *Moreover,* $c\in C^{2,1}(\bar{D}_T)$, $c_{xt}\in C(\bar{D}_T\setminus\{0,0\})$, $s\in C^2[0,T]$.

Proof. This is a straightforward consequence of Theorem 3.1 and of Banach's fixed point theorem. The regularity properties of c and s follow from Proposition 2.3 and the definition of X. □

A monotone dependence lemma will be useful in proving the uniqueness theorem. Let c_i, s_i, $i=1,2$, solve the problems

$$
c_{ixx}-c_{it}=0, \qquad 0<x<s_i(t), \quad t_i<t<T, \tag{3.8}
$$

with initial conditions $s_i(t_i)=0$ and satisfying boundary conditions (1.3)–(1.5) in the time intervals (t_i,T). We have

LEMMA 3.3. *If* $t_1<t_2$, *then*

$$
s_1(t)>s_2(t), \qquad t_2<t<T. \tag{3.9}
$$

Proof. Note that the transformation

$$
u(x,t)=-\int_x^{s(t)}\left[c(y,t)+q\right]dy, \tag{3.10}
$$

carries (1.1)–(1.5) into the following Stefan-like problem:

$$
u_{xx}-u_t=0 \qquad \text{in } D_T, \tag{3.11}
$$

$$
s(0)=0, \tag{3.12}
$$

$$
u_x(0,t)=1+q, \qquad 0<t<T, \tag{3.13}
$$

$$
u(s(t),t)=0, \qquad 0<t<T, \tag{3.14}
$$

$$
u_x(s(t),t)=\Phi(\dot{s}(t))+q, \qquad 0<t<T. \tag{3.15}
$$

Consider the function $u_i(x,t)$ obtained from $c_i(x,t)$ by means of (3.10). Assume that there exists a first time t_0 such that $s_1(t_0)=s_2(t_0)$ and hence

$$\dot{s}_1(t_0)\leqq\dot{s}_2(t_0). \tag{3.16}$$

Standard use of the maximum principle applied to the difference u_1-u_2 leads to a contradiction of (3.16). □

Now we can prove uniqueness.

THEOREM 3.4. *Problem* (P) *cannot have two distinct solutions with the same* T.

Proof. Let (T,s,c) and (T,σ,γ) be two solutions and consider

$$s_n^{\pm}(t)=s(t\pm 1/n),\qquad c_n^{\pm}(x,t)=c(x,t\pm 1/n).$$

Note that s_n^+, c_n^+ and s_n^-, c_n^- are solutions corresponding to initial data $s_n^+(-1/n)=0$ and $s_n^-(1/n)=0$, respectively. According to Lemma 3.3

$$s_n^-(t)<\sigma(t)<s_n^+(t),\qquad 1/n<t<T-1/n.$$

Letting n tend to infinity concludes the proof to the theorem. □

Remark 3.5. As a consequence of Theorems 3.2 and 3.4, we have that for any solution of problem (P) a T_0 can be found such that $c\in C^{2,1}(\bar{D}_{T_0})$, $c_{xt}\in C(\bar{D}_T LT_0\setminus\{0,0\})$, $s\in C^2[0,T_0]$.

4. Regularity, convexity, global existence, continuous dependence. Before proving global existence, let us perform an a priori analysis on the solutions of problem (P).

PROPOSITION 4.1. *Assume* s,c *solve problem* (P) *for a given* $T<+\infty$. *Then, there exists* $c_0=c_0(T)>0$ *such that*

$$c_0<c(x,t)<1,\qquad 0<x\leqq s(t),\quad 0<t<T, \tag{4.1}$$

$$0<f(c_0)\equiv\dot{r}_0\leqq\dot{s}(t)\leqq f(1),\qquad 0\leqq t<T, \tag{4.2}$$

$$0>c_x(x,t)>-f(1)[q+1],\quad \text{in } D_T. \tag{4.3}$$

Proof. If $c(x,t)$ attains the value 0 (necessarily at $x=s(t)$) for the first time at some point $(s(t_0),t_0)$, one would have $c_x(s(t_0),t_0)=0$, contradicting the boundary point principle. This yields $c\geqq c_0$, $\dot{s}\geqq\dot{r}_0$. The last inequalities in (4.1), (4.2) follow as in the proof of Proposition 2.2. The same is true for inequality (4.3). □

Now we prove

THEOREM 4.2. *Let* (T,s,c) *be a solution to problem* (P). *Then* $s\in C^2[0,T]$. *Moreover, if* $f\in C^\infty(0,1]$ *then* $s\in C^\infty(0,T)$.

Proof. This result can be proved either directly, or by applying the iterative technique introduced in [9] to the equivalent problem (3.11)–(3.15), with (3.15) written more conveniently as $\dot{s}(t)=f[u_x(s(t),t)-q]$. Following the second way, start from $t_0\in(0,T)$ and conclude step by step that the derivatives up to u_{xt} are Hölder continuous on $x=s(t)$ for $t\in(t_0,T)$]. At this point the continuity of f' ensures the continuity of $\ddot{s}$. If $f\in C^\infty(0,1]$ the same procedure can be iterated indefinitely, thus concluding that $s\in C^\infty(0,T]$. The continuity of $\ddot{s}(t)$ at $t=0$ is already known (see Remark 3.5). □

At this point, we can prove the convexity of the free boundary.

THEOREM 4.3. *Assume* (T,s,c) *solve Problem* (P). *Then*

$$\ddot{s}(t)<0,\qquad 0\leqq t\leqq T. \tag{4.4}$$

Proof. Theorem 4.2 yields the continuity of c_t in $\overline{D}_T$ and the continuity of c_{xt} in $\overline{D}_T \setminus \{0,0\}$. Hence the function

$$v(x,t) = [\ln(c+q)]_{xx}$$

is continuous in $\overline{D}_T$ and v_x is continuous in $\overline{D}_T \setminus \{0,0\}$. Note that

$$v_{xx} + 2[\ln(c+q)]_x v_x + 2v^2 - v_t = 0 \quad \text{in } D_{T_0} \tag{4.5}$$

and that

$$v(0,t) = -[c_x/(c+q)]^2\big|_{x=0} < 0,$$
$$v(s(t),t) = \ddot{s}(t)\Phi'(\dot{s})/(q+\Phi(\dot{s})).$$

Owing to the continuity of $\ddot{s}(t)$ for $t=0$, we can assert that $\ddot{s}(t)$, and consequently $v(s(t),t)$, is negative in some interval $[0,t_0)$ (recall that $\ddot{s}(0) = -f^1(1)f^2(1)(1+q) < 0$).

The maximum principle (e.g in the form of [6, Thm. 5, p. 39]) applied with some care to (4.5) implies that if $v(s(t),t)$ vanishes for the first time at some $t_0 > 0$, then $v(x,t) < 0$ in D_{t_0}. Therefore equation (4.5) is such that the boundary point principle (see e.g. [6, Thm. 14, p. 49]) can be applied, yielding $v_x(s(t_0),t_0) > 0$. However, $v_x(s(t),t) = \dot{s}(t)\Phi'(\dot{s})\ddot{s}(t)/(q+\Phi(\dot{s}))$, contradicting $\ddot{s}(t_0) = 0$. □

COROLLARY 4.4. *Under the same assumptions*

$$c_t(x,t) > 0, \qquad 0 < x \leqq s(t), \quad 0 < t < T_0, \tag{4.6}$$

$$c_{xt}(0,t) > 0, \qquad 0 < t < T_0. \tag{4.7}$$

Proof. Indeed, as in the proof of Proposition 2.3, we write the problem solved by $w = c_t$ for which we have the estimate (2.14). But, (4.4) implies $F(t) \geqq 0$, whence $c_t \geqq 0$ in $\overline{D}_T$. At this point (4.6) follows by the strong maximum principle. Since $c_t(0,t) = 0$, (4.7) follows by the boundary point principle. □

COROLLARY 4.5. *Under the same assumptions*

$$\ddot{s}(t) \geqq -[f(1)]^2 \cdot [q+1] \cdot \max_{\eta \in [c_0,1]} f'(\eta), \qquad 0 < t < T_0. \tag{4.8}$$

Proof. Since

$$\ddot{s}(t) = f'(c(s(t),t))[c_x \dot{s} + c_t]_{x=s(t)}, \qquad 0 < t < T_0, \tag{4.9}$$

(4.8) follows from (4.2), (4.3), (4.6). □

Remark 4.6. Estimate (4.6) can be obtained independently of (4.4), by simply noting that the use of both the free boundary conditions enables us to write the boundary problem for $w = c_t$ replacing (2.13) by

$$[w\dot{s} + w_x]_{x=s(t)} = -\frac{d}{dt}[f(c)(q+c)]_{s=s(t)},$$

i.e.

$$[w_x + (\dot{s} + f + (1+c)f')w]_{x=s(t)} = -[c_x \dot{s}\{f + (q+c)f'\}]_{x=s(t)}.$$

Since the r.h.s. is nonnegative, (4.6) follows from the maximum principle.

THEOREM 4.7. *Problem* (P) *admits a solution for arbitrary* $T > 0$.

Proof. Assume there exists a $T^* > 0$ such that the solution (whose local existence was proved in Theorem 3.2) cannot be continued beyond T^*. We have that

$$\lim_{t \to T^*_-} s(t), \qquad \lim_{t \to T^*_-} \dot{s}(t)$$

both exist, because of the monotonicity of s and $\dot{s}$ (see Proposition 4.1 and Theorems 4.2, 4.3). Consider the free boundary problem

$$u_{xx}-u_t=0, \qquad 0<x<s(t), \quad T^*<t<\hat{T},$$

with boundary data given by (3.13)–(3.15) (in the time interval $(T^*,\hat{T})$) and "initial" data given by the limits for $t\to T^*-$ of $s(t)$ and of $\int_x^{s(t)}[c(\xi,t)+q]\,d\xi$. This problem has a unique solution (for suitable $\hat{T}>T^*$), provided $\Phi'(\dot{s}(T^*))\neq 0$ (see [5]). This fact is guaranteed by (4.2) and by assumption (F). Hence the theorem. □

Finally, we want to investigate the dependence of the solution on the data. Assume f_1, f_2 (both satisfying assumption (F)) and q_1, q_2 (both satisfying $0\leqq q_i\leqq Q$) are data for problem (1.1)–(1.5) and let $s_i(t)$, $c_i(x,t)$ be the corresponding solutions in a given interval $(0,T)$. We know that s_1 and s_2 are the fixed points of operators $\mathscr{C}_1$ and $\mathscr{C}_2$ mapping the same set $X(\overline{K},\overline{T},\overline{\gamma})$ into itself.

Now, the contractive character of operator $\mathscr{C}(f,q)$ does not depend on the pair f_i, q_i but only on Q, $\dot{r}_0$.

Moreover, $\mathscr{C}$ depends continuously on $|q_1-q_2|+\|f_1-f_2\|_{C^1(0,\overline{T})}$. Beyond $\overline{T}$, the continuous dependence follows from the results of [5], applied to the problem solved by $u(x,t)$ defined according to (4.15). Thus we have proved:

THEOREM 4.8. *The solution of problem* (P) *depends continuously upon the data* q,f.

5. Asymptotic estimates. Let $s(t)$, $c(x,t)$ solve problem (P) for any $T>0$. Then, Green's identity

$$0=\int\int_{D_t} x(c_{xx}-c_t)\,dx\,d\tau=\oint_{\partial D_t}\left[(xc_x-c)\,d\tau+xc\,dx\right], \qquad t>0,$$

gives

$$\frac{1}{2}qs^2(t)-t+\int_0^{s(t)} xc(x,t)\,dx+\int_0^t c(s(\tau),\tau)d\tau=0, \qquad t>0. \tag{5.1}$$

Moreover we have

PROPOSITION 5.1. *The following estimates hold in* D_T, $\forall T>0$.

$$c(x,t)\leqq 1+x[\Phi(\dot{s}(t))-1]/s(t), \tag{5.2}$$

$$c(x,t)\geqq \Phi(\dot{s}(t))-\dot{s}(t)[\Phi(\dot{s}(t))+q](x-s(t)). \tag{5.3}$$

Proof. Both inequalities follow from the convexity of the curve, representing $c(x,t)$ as a function of x, for any $t>0$ (see Theorem 4.3). □

Now, we prove

THEOREM 5.2.

$$\lim_{t\to+\infty} s(t)=+\infty, \tag{5.4}$$

$$\lim_{t\to+\infty} \dot{s}(t)=0. \tag{5.5}$$

Proof. The existence of both limits is a consequence of the monotonicity of $s,\dot{s}$. Also $c(s(t),t)$ and $c(x,t)$ have limits for $t\to+\infty (c_t>0, dc(s(t),t)/dt<0)$.

Let (5.4) be false; then $-t$ and possibly $\int_0^t c(s(\tau),\tau)\,d\tau$ are the only terms in (5.1) going to infinity as $t\to+\infty$. But $c(s(t),t)$ tends monotonically to zero like $\dot{s}(t)$, and hence the two terms are of different order. This proves (5.4).

To prove (5.5), just note that the compatibility of (5.2) and (5.3) in $x=0$ requires

$$\Phi(\dot{s}(t))[1+s(t)\dot{s}(t)]+qs(t)\dot{s}(t)\leqq 1. \tag{5.6}$$

Note that this argument applies also when $q=0$. □

THEOREM 5.3. *If* $q>0$, *then*

$$s(t)\leqq\sqrt{\frac{2t}{q}} \quad \text{as } t\to+\infty. \tag{5.7}$$

If $q\geqq 0$, *then*

$$s(t)\geqq\sqrt{\frac{2}{q+1/3}}\sqrt{t}\,[1-\varepsilon^2(t)], \tag{5.8}$$

where $\lim_{t\to+\infty}\varepsilon^2(t)=0$.

Proof. To prove (5.7), it suffices to use the positivity of c and (5.1). To prove (5.8), we note that (5.2) implies

$$\int_0^{s(t)} xc(x,t)\,dx\leqq\frac{s^2(t)}{2}+\frac{s^2(t)}{3}[\Phi(\dot{s}(t))-1].$$

Hence, from (5.1) we have

$$\left[\frac{1}{3}\Phi(\dot{s}(t))+\frac{1}{2}q+\frac{1}{6}\right]s^2(t)\geqq t\left[1-\frac{1}{t}\int_0^t\Phi(\dot{s}(\tau))\,d\tau\right]. \tag{5.9}$$

But $\Phi(\dot{s}(t))$ and $(1/t)\int_0^t\Phi(\dot{s}(\tau))\,d\tau$ tend to zero as $t\to+\infty$; hence (5.8) follows. □

Now, we want to investigate the case $q=0$. To this aim, we need to know the behaviour of $\Phi(z)$ near $z=0$. To be specific, we will assume that

$$\Phi(z)=\alpha z^{1/m} \quad \text{for some } \alpha,\ m>0. \tag{5.10}$$

We can prove

THEOREM 5.4. *Let* $q=0$ *and* (5.10) *hold. Then*

$$s(t)\leqq\left(\frac{2}{\alpha}\right)^{m/(2m+1)}\left(\frac{2m+1}{m+1}\right)^{1/(2m+1)}t^{(m+1)/(2m+1)}. \tag{5.11}$$

Proof. From (5.1) we have

$$\int_0^{s(t)} xc(x,t)\,dx-t<0.$$

Since $c(x,t)>\Phi(\dot{s}(t))$, this implies

$$\Phi(\dot{s}(t))s^2(t)\leqq 2t. \tag{5.12}$$

Using (5.9), we have

$$\alpha^m s^{2m}(t)\dot{s}(t)\leqq 2^m t^m,$$

whence (5.11) follows by integration. □

Numerical calculations (see next section) show that the right-hand side of (5.8) is a good approximation of $s(t)$ for t sufficiently large. In this respect it is of some interest to estimate the term $\varepsilon^2(t)$ in (5.8). To this end we can utilize the inequality

$$\dot{s}(t)\leqq\min\left[(qs(t))^{-1},(\alpha s(t))^{-m/(1+m)}\right], \tag{5.13}$$

following from (5.6), (5.10). From (5.9) it follows that

$$(5.14)\qquad s(t)\geqq\sqrt{\frac{2}{q+1/3}}\sqrt{t}\left\{1-\frac{1}{t}\int_0^t\Phi(\dot{s}(\tau))\,d\tau-2\Phi(\dot{s}(t))/(3q+1)\right\}^{1/2}.$$

For some fixed $t_0>0$ the quantity in braces is less than 1 and approaches 1 for t_0 large. Let us write it $1-\delta$ and use the consequent estimate in (5.13) to get upper bounds for $\dot{s}(t)$. Now use such bounds in turn in (5.14) to conclude the following

(i) If $q=0$, then

$$\varepsilon^2(t)\cong\alpha\frac{m+2}{2m+1}\left[6\alpha^2(1-\delta)t\right]^{-1/2(1+m)}.$$

(ii) If $q>0$, then

$$\varepsilon^2(t)\cong\alpha\left(\frac{m}{2m-1}+\frac{1}{3q+1}\right)q^{-1/m}\left(\frac{q+1/3}{2(1-\delta)}\right)^{1/2m}t^{-1/2m},\qquad m>\tfrac{1}{2},$$

for $m=\frac{1}{2}$, $\varepsilon^2(t)$ behaves like $(\log t)/t$ and for $m<\frac{1}{2}$ like $1/t$.

6. The numerical method. In this section will be shown the convergence of a numerical scheme based on the method introduced in [8] for one-dimensional parabolic free boundary problems with arbitrary implicit or explicit free boundary conditions.

In this method the continuous problem is time discretized and solved at successive time levels as a sequence of free boundary problems for ordinary differential equations. Specifically, at the time level $t=t_n$ with $t_n-t_{n-1}=\Delta t$ the solution $\{C_n(x),S_n\}$ is computed as the exact solution of the discretized equations

$$(6.1)\qquad C_n''-\frac{1}{\Delta t}(C_n-C_{n-1})=0,\qquad 0<x<S_n,$$

$$(6.2)\qquad C_n(0)=1,$$

$$(6.3)\qquad \frac{S_n-S_{n-1}}{\Delta t}=f(C_n(S_n)),\qquad S_0=S(0)=0,$$

$$(6.4)\qquad C_n'(S_n)=-\frac{S_n-S_{n-1}}{\Delta t}(q+C_n(S_n)).$$

In (6.1) the function $C_{n-1}(x)$ is supposed to be defined over $[0,+\infty)$, and S_{n-1} is supposed to be known as well. The free boundary problem (6.1)–(6.4) is conveniently solved with the method of invariant imbedding (sweep method). We write (6.1) as a first order system over $(0,S_n)$

$$(6.5)\qquad C_n'=V_n,$$

$$(6.6)\qquad V_n'=\frac{1}{\Delta t}(C_n-C_{n-1})$$

and exploit the observation that C_n and V_n are related through the Riccati transformation

$$(6.7)\qquad C_n(x)=R(x)V_n(x)+W_n(x),$$

where

$$(6.8)\qquad R'=1-\frac{1}{\Delta t}R^2,\qquad R(0)=0,$$

$$(6.9)\qquad W_n'=-\frac{R(x)}{\Delta t}(W_n-C_{n-1}(x)),\qquad W_n(0)=1.$$

The functions R and W are solutions of well defined initial value problems and may be considered available. The free boundary S_n is determined such that the triple C_n, V_n, S_n simultaneously satisfies (6.3), (6.4), and (6.7). Elimination of C_n and V_n from (6.4) and (6.7) shows that S_n must be a root of the scalar equation

$$\sigma_n(x) \equiv (x - S_{n-1})/\Delta t - f\left[(W_n(x) - qR(x)(x - S_{n-1})/\Delta t)(1 + R(x)(x - S_{n-1})/\Delta t)^{-1}\right] = 0 \tag{6.10}$$

(extend f as an odd function). Given S_n, we set

$$C_n(S_n) = \frac{W_n(S_n) - R(S_n)\dot{S}_n q}{1 + R(S_n)\dot{S}}, \tag{6.11}$$

so that

$$\dot{S}_n \equiv \frac{S_n - S_{n-1}}{\Delta t} = f(C_n(S_n)), \tag{6.12}$$

and

$$C_n'(S_n) \equiv V_n(S_n) = -\frac{S_n(W_n(S_n) + q)}{1 + R(S_n)\dot{S}_n}. \tag{6.13}$$

Thus, the triple $\{C_n(S_n), V_n(S_n), S_n\}$ is an exact solution of (6.3), (6.4) and (6.7). We remark that depending on Δt and q the functional $\sigma_n(x)$ may have a root smaller than S_{n-1}. Such a root would correspond to a negative concentration $C_n(S_n)$ and is not admissible. We shall therefore agree to choose for S_n the smallest root of $\sigma_n(x) = 0$ on (S_{n-1}, ∞). Such a root will be shown to exist.

Once S_n has been determined, one can find V_n by integrating backward over $[0, S_n)$ the reverse sweep equation

$$V_n' = \frac{1}{\Delta t}\left[R(x)V_n + W_n(x) - C_{n-1}(x)\right] \tag{6.14}$$

with $V_n(S_n)$ given by (6.13). The concentration $C_n(x)$ at time level T_n is obtained from (6.7). Finally, $C_n(x)$ is extended over $[S_n, \infty)$ as C^1 linear function. For the initial concentration we shall use

$$C_0(x) = 1 + C_0' x = 1 - f(1)(q+1)x.$$

Assuming f satisfies (F), §1, we will derive some estimates of the solution $\{C_n, S_n\}$ to (6.1)–(6.4). Such estimates correspond to the bounds obtained in previous sections for the solution of the continuous problem.

LEMMA 6.1. *There exists a solution S_n of* (6.10) *on* (S_{n-1}, ∞) *and* C_n *satisfies* $0 < C_n \leqq 1$ *on* $[0, S_n]$ *and* $C_n' < 0$ *on* $[0, \infty)$.

Proof. We note that $C_0(S_0) = 1$ and $C_0' < 0$. Assume next that $C_{n-1}(S_{n-1}) > 0$ and $C_{n-1}' < 0$. We observe from $R(x) > 0$ on $(0, \infty)$ and

$$(W_n - C_{n-1})' = \frac{R(x)}{\Delta t}(W_n - C_{n-1}) - C_{n-1}',$$

$(W_n - C_{n-1})(0) = 0$ that $W > C_{n-1}$ on $(0, \infty)$ and $W_n' \leqq 0$ on $[0, \infty)$. Moreover $\sigma_n(S_{n-1}) = -f(W_n(S_{n-1})) \leqq -f(C_{n-1}(S_{n-1})) < 0$ and

$$\lim_{x \to \infty} \sigma_n(x) = \infty.$$

Thus there must be a point $S_n > S_{n-1}$ where $\sigma_n(S_n)=0$, $C_n(S_n)>0$ and $C_n'(S_n)<0$. The linear equation (6.14) now assures that $V_n<0$ and that $0<C_n\leqq 1$ on $[0,S_n]$. Hence the lemma is true for all n. □

We remark that the monotonicity of $R(x)$ and $W_n(x)$ assures that $\sigma_n'>0$ (S_{n-1},∞). We also note that

$$0<S_n-S_{n-1}=f(C_n(S_n))\Delta t\leqq f(1)\Delta t. \tag{6.15}$$

LEMMA 6.2. $-f(1)(q+1)\leqq C_n'(x)<0$ *on* $[0,\infty)$.

Proof. The upper bound is guaranteed by Lemma 6.1. The lower bound is obtained by applying the maximum principle to

$$\left(C_n'\right)''-\frac{1}{\Delta t}C_n'=-\frac{1}{\Delta t}C_{n-1}',$$
$$C_n''(0)=0,\qquad C_n'(S_n)=-f(C_n(S_n))(q+C_n(S_n))$$

from which it follows that

$$\min C_n'\geqq\min\left\{\min C_{n-1}', C_n'(S_n)\right\}.$$

Since $C_n'(S_n)\geqq -f(1)(q+1)$ and $C_0'=-f(1)(q+1)$ the lemma is proved. □

LEMMA 6.3. $0\leqq C_n-C_{n-1}\leqq f(1)^2(q+1)\Delta t$ *on* $[0,S_n]$ *for* $n=1,2,\cdots$.

Proof. We observe from Lemma 6.2 that $(C_1-C_0)'\geqq 0$ so that $C_1-C_0\geqq 0$. Let us assume that $C_{n-1}-C_{n-2}\geqq 0$ and consider the problem satisfied by C_n-C_{n-1} on $(0,S_n)$:

$$(C_n-C_{n-1})''-\frac{1}{\Delta t}(C_n-C_{n-1})=\begin{cases}-\dfrac{1}{t}(C_{n-1}-C_{n-2}), & x\in(0,S_{n-1}),\\ 0, & x\in(S_{n-1},S_n)\end{cases}$$
$$(C_n-C_{n-1})(0)=0,$$
$$(C_n-C_{n-1})'(S_n)=-f(C_n)(q+C_n)\big|_{S_n}+f(C_{n-1})(q+C_{n-1})\big|_{S_{n-1}}.$$

The maximum principle assures that its solution has no negative minimum on the half open interval $[0,s_n)$ and no positive maximum on (S_{n-1},S_n). At a positive maximum on $(0,S_{n-1}]$ the maximum principle yields $C_n-C_{n-1}\leqq\max[C_{n-1}-C_{n-2}]$. Let us now consider the remaining case when C_n-C_{n-1} has an extremum at S_n. The boundary condition can be rewritten with the mean value theorem as

$$(C_n-C_{n-1})'(S_n)=-\left[f'(\xi)(q+\xi)+f(\xi)\right]\cdot\left[(C_n-C_{n-1})(S_n)+C_{n-1}(S_n)-C_{n-1}(S_{n-1})\right]$$

for some $\xi>0$. Since $C_{n-1}(S_n)-C_{n-1}(S_{n-1})<0$ by Lemma 6.2 this boundary condition rules out a negative minimum at S_n. Hence $C_n-C_{n-1}\geqq 0$ on $[0,S_n]$. Suppose finally that C_n-C_{n-1} assumes its maximum at S_n. Then $(C_n-C_{n-1})'(S_n)\geqq 0$ so that the boundary condition implies that

$$0\leqq(C_n-C_{n-1})(S_n)\leqq C_{n-1}(S_{n-1})-C_{n-1}(S_n)=C_{n-1}'(\eta)(S_{n-1}-S_n)$$
$$\leqq f(1)^2(q+1)\Delta t\quad\text{for some }\eta\in(S_{n-1},S_n).$$

Since this argument already applies to C_1-C_0 on $[0,S_1]$ the lemma follows by induction. □

We note from

$$C_n(S_n)-C_{n-1}(S_{n-1})=C_n(S_n)-C_{n-1}(S_n)+C_{n-1}(S_n)-C_{n-1}(S_{n-1})$$

that

$$C_{n-1}(S_n)-C_{n-1}(S_{n-1})\leqq C_n(S_n)-C_{n-1}(S_{n-1})\leqq C_n(S_n)-C_{n-1}(S_n)$$

so that

$$|C_n(S_n)-C_{n-1}(S_{n-1})|\leqq f(1)^2(q+1)\Delta t. \tag{6.16}$$

LEMMA 6.4. $|(C_n-C_{n-1})'|\leqq K\Delta t$ *for* $n=1,2,\cdots$.

Proof. Let $Q_n=C_n'$ then it follows from Lemma 6.3 and $f(1)^2(q+1)\geqq (Q_n-Q_{n-1})'=(1/\Delta t)(C_n-C_{n-1})\geqq 0$ on $[S_{n-1},S_n)$ that Q_n-Q_{n-1} is increasing on $[S_{n-1},S_n]$. We also note that as in Lemma 6.3

$$(Q_n-Q_{n-1})(S_n)=-[f'(\xi)(q+\xi)+f(\xi)](C_n(S_n)-C_{n-1}(S_{n-1}))$$

so that by (6.16)

$$|(Q_n-Q_{n-1}(S_n)|\leqq f(1)^2(q+1)\Delta t\cdot\max_{\xi\in[0,1]}[f'(\xi)(q+\xi)+f(\xi)]\equiv K_1\Delta t$$

and

$$|(Q_n-Q_{n-1})(S_n)|\leqq K_1\Delta t+f(1)^3(q+1)\Delta t\equiv K\Delta t.$$

These estimates imply in particular that

$$|Q_1-Q_0|\leqq K\Delta t.$$

Finally, we apply the maximum principle to

$$(Q_n-Q_{n-1})''-\frac{1}{\Delta t}(Q_n-Q_{n-1})=-\frac{1}{\Delta t}(Q_{n-1}-Q_{n-2})\quad\text{on } (0,S_{n-1}),$$
$$(Q_n-Q_{n-1})'(0)=0$$

to conclude $|Q_n-Q_{n-1}|\leqq\max\{|(Q_n-Q_{n-1})(S_{n-1})|,\max|Q_{n-1}-Q_{n-2}|\}$ and hence that $|Q_n-Q_{n-1}|\leqq K\Delta t$ for all n. □

We observe that it is a consequence of these lemmas that

$$|C_n-C_{n-1}|\leqq\tilde{K}\Delta t$$

for some $\tilde{K}>0$ on compact subsets of $[0,\infty)$. In fact, suppose $\tilde{x}>S_n$ then

$$C_n(\tilde{x})-C_{n-1}(\tilde{x})=C_n(S_n)-C_{n-1}(S_n)+[C_n'(S_n)-C_{n-1}'(S_n)](\tilde{x}-S_n).$$

The desired inequality now follows from Lemmas 6.3 and 6.4. For arbitrary but fixed T let us define the interpolating functions $S_N(t)$ on $[0,T]$ and $C_N(x,t)$ on $[0,f(1)]\times[0,T]$ given by

$$S_N(t)=S_{n-1}+\frac{S_n-S_{n-1}}{\Delta t}(t-t_{n-1}),\qquad t\in[t_{n-1},t_n),$$
$$C_N(x,t)=C_{n-1}(x)+\frac{C_n(x)-C_{n-1}(x)}{\Delta t}(t-t_{n-1}),\qquad t\in[t_{n-1},t_n)$$

where $\Delta t = T/N$. It follows from (6.15) and from Lemmas 6.2, 6.3 and 6.4 that both S_N, C_N and $\partial C_N/\partial x$ are uniformly bounded and Lipschitz continuous on $[0,T]$ and $[0,f(1)T]\times[0,T]$. By the Ascoli–Arzela theorem a subsequence of $\{S_N, C_N\}$ will converge uniformly to Lipschitz continuous limit functions $\{s(t), c(x,t)\}$ defined for $0 \leqq t \leqq T$ and $0 \leqq x \leqq f(1)T$. Moreover, $(\partial c/\partial x)(x,t)$ is Lipschitz continuous and the uniform limit of $\partial C_N/\partial x$.

LEMMA 6.5. *The limit function $s(t)$ has a Lipschitz continuous derivative.*

Proof. Let $t \in (0,T]$ be arbitrary and set $\Delta t = t/N$. Define

$$Q(t) = \int_0^t f(c(s(\tau),\tau))\, d\tau,$$

then dQ/dt is Lipschitz continuous because f is continuously differentiable and c and s themselves are Lipschitz continuous. Let $\{C_n, S_N\}$ denote a convergent subsequence. Then by construction

$$S_N(t) = \Delta t \sum_{i=1}^{N} (S_i - S_{i-1}) = \Delta t \sum_{i=1}^{N} f(C_i(S_i)) = \Delta t \sum_{i=1}^{N} f(C_N(S_N(t_i),t_i)).$$

It now follows from

$$Q(t) - S_N(t) = \left[\int_0^t f(c(s(\tau),\tau))\, d\tau - \Delta t \sum_{i=1}^{N} f(c(s(t_i),t_i))\right]$$

$$+ \Delta t \sum_{i=1}^{N} [f(c(s(t_i),t_i)) - f(C_N(S_N(t_i),t_i))]$$

that $\lim_{N\to\infty}[Q(t) - S_N(t)] = 0$. Indeed, the first bracket on the right is the difference between the integral and its Riemann sum approximation and hence vanishes as $\Delta t \to 0$. The second bracket vanishes because C_N and S_N converge uniformly to Lipschitz continuous limit functions.

We have thus shown that the limit functions s and c satisfy

$$\dot{s}(t) = f(c(s(t),t)).$$

It also is readily seen that

$$\frac{\partial c}{\partial x}(s(t),t) = -f(c(s(t),t))(q + c(s(t),t))$$

since

$$\frac{\partial C_N}{\partial x}(S_N(t),t) = -f(C_N(S_N(t),t))(q + C_N(S_N(t),t))$$

at each point of the partition $\{t_i\}$ associated with N. Since these points are dense in $[0,T]$ as $N \to \infty$ and the convergence of $\partial C_N/\partial x$, C_N and S_N is uniform, it follows immediately that the limit functions c, s also satisfy the second boundary condition (1.5).

To conclude that the whole sequence of interpolating functions converges, it is enough to recall the uniqueness of the solution to problem (P), see Theorem 3.4.

7. A numerical example. The algorithm of the preceding section is used to examine numerically the asymptotic estimates (5.7), (5.8) for the special case of $m = 2$ and $q = 1$. All calculations are carried out with the research code of [8] as follows. For a

given interval $[0, X]$ a variable but time independent mesh with grid points $\{x_k\}_{k=0}^{N}$ is imposed with $x_i = X(i/N)^2$ so that the mesh points cluster near the origin where the free boundary moves fast. The time step Δt is variable and increases with time. For the Riccati equation (6.8) the analytic solution

$$R(x, \Delta t) = \sqrt{\Delta t} \tanh \frac{x}{\sqrt{\Delta t}}$$

is used, while the linear equation (6.9) is integrated with the trapezoidal rule. The function $\sigma_n(x)$ is evaluated at successive mesh points of the grid beyond S_{n-1} until it changes sign between, say, x_l and x_{l+1}. The free boundary S_n is now determined as the root of the quadratic interpolant through $\sigma(x_{l-1})$, $\sigma(x_l)$ and $\sigma(x_{l+1})$. The equation (6.14) is also integrated with the trapezoidal rule, first from S_n to x_l and then backward over the fixed mesh. The numerical methods for these equations are known to converge as $\Delta x \to 0$. The results obtained are reasonably stable with respect to changes in the mesh. However, if run times are to be minimized, then a Crank–Nicolson time discretization and a higher order Adams–Moulton space integration are suggested.

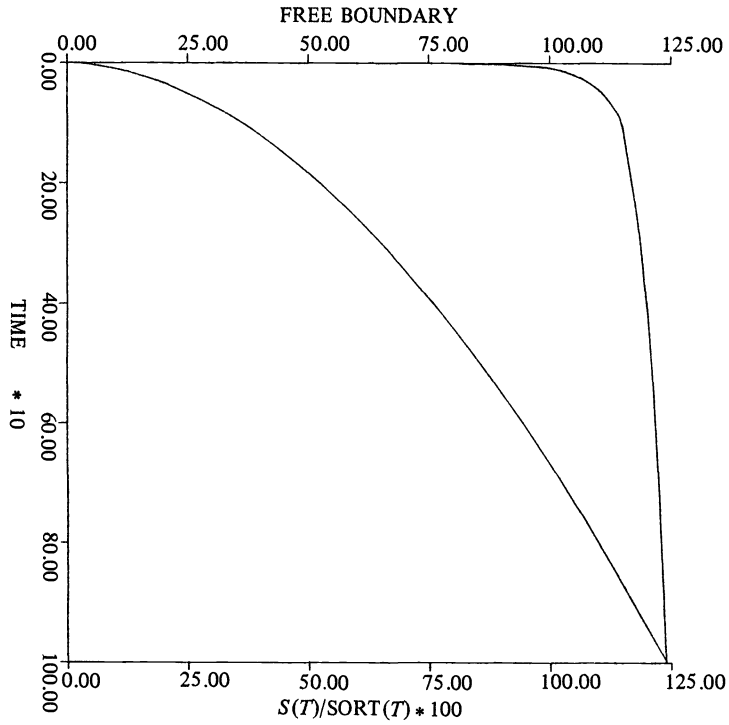


FIG. 1. *Plot of the free boundary and its asymptotic behavior for* $q=1$, $m=2$. $X=140$, $N=6000$ *with* 100 *equal time steps for each of the intervals* $[0,1]$, $[1,10]$, $[10,100]$, $[100,10000]$, *and* $[1000,10000]$. *Total execution time on the Cyber* 855 *was* 200*s*.

Figure 1 shows a plot of the free boundary $s(t)$ for $t \in [0, 10000]$ and $\phi(t) = s(t)/\sqrt{t}$. The computed values satisfy the inequalities

$$\sqrt{1.5}\,(1-\varepsilon^2(10000)) \leqq \sqrt{1.5} \equiv 1.225 \leqq \phi(10000) = 1.24 \leqq \sqrt{2} = 1.41,$$

as well as $\phi'(10000) \approx 4.18 \cdot 10^{-6}$ and $s'(10000) \approx 6.62 \cdot 10^{-3}$. Thus the lower bound (5.8) appears to be quite good.

REFERENCES

[1] G. ASTARITA, *A class of free boundary problems arising in the analysis of transport phenomena in polymers*, in Free Boundary Problems: Theory and Applications, A. Fasano and M. Primicerio, eds., Vol. II, Pitman, London, 1983, pp. 602–612.

[2] G. ASTARITA AND G. C. SARTI, *A class of mathematical models for sorption of swelling solvents in glassy polymers*, Polym. Eng. Sci., 18 (1978), pp. 388–395.

[3] T. BRUNO AND G. MONTAGNARO, *Su un problema di frontiera mobile in cui il fronte è una linea di discontinuita*, Rend. Accad. Sci. Fis. Mat., IV, 49 (1982), pp. 51–71.

[4] C. COSTOLI AND G. C. SARTI, *Diffusion and localized resistances in glassy polymers*, Polym. Eng. Sci., 22 (1982), pp. 1018–1026.

[5] A. FASANO AND M. PRIMICERIO, *Free boundary problems for nonlinear parabolic equations with nonlinear free boundary conditions*, J. Math. Anal. Appl., 72 (1979), pp. 247–273.

[6] A. FRIEDMAN, *Partial Differential Equations of Parabolic Type*, Prentice-Hall, Englewood Cliffs, NJ, 1964.

[7] O. A. LADYŽENSKAYA, V. A. SOLONNIKOV AND N. N. URAL'CEVA, *Linear and quasilinear equations of parabolic type*, American Mathematical Society Translations 23, Providence, RI, 1968.

[8] G. H. MEYER, *One-dimensional parabolic free boundary problems*, SIAM Rev., 19 (1977), pp. 17–34.

[9] D. G. SCHAEFFER, *A new proof of the infinite differentiability of the free boundary in the Stefan problem*, J. Differential Equations, 20 (1976), pp. 266–269.

[10] YIH-O TU, *A multi-phase Stefan problem describing the swelling and dissolution of glassy polymer*, Quart. Appl. Math., 35 (1977), pp. 269–285.

SIAM J. MATH. ANAL.
Vol. 17, No. 4, July 1986

014

ON FUNCTIONS REPRESENTABLE AS A SUPREMUM OF A FAMILY OF SMOOTH FUNCTIONS II*

Y. YOMDIN†

Abstract. The classes S_m^q of functions $f(x)$, representable as $\sup_t h(x,t)$, where t is an m-dimensional parameter and h is a C^q-smooth function of x and t, are studied. Considering the "massiveness" of the sets S_m^q in appropriate functional spaces, we show that these classes really differ for different q and m.

Studying geometric invariants of maximum functions, related to the critical values of smooth mappings involved, we give explicit examples of functions in S_m^q which are not representable as the maximum of k times differentiable families when k is sufficiently large.

1. Introduction. Functions, represented as a supremum of families of a certain type, arise naturally in many questions of analysis and optimization (see e.g. [1], [5], [6], [13] and many others). In [7], [9], [12] the class $H(D)$ has been considered of functions f on the domain $D \subset R^n$, representable as $f(x) = \sup_{h \in Q} h(x)$, $x \in D$, where Q is a bounded in a C^2-norm family of twice differentiable functions on D. The functions $f \in H(D)$ have many nice properties, both geometric and analytic. One important point is that the consideration of families of C^k-smooth functions, bounded in C^k-norm, for $k > 2$, does not restrict the class H. In fact, it is shown in [7], [9] that $H(D)$ can be described as the class of all f, representable as $f(x) = \sup_{p \in P} p(x)$, where P is a bounded subset in the space of all the quadratic polynomials on R^n.

Another important class of maximum functions appears when we assume that the family Q is smoothly parametrized.

Let $S_m^q(D)$ denote the set of functions r, representable as $f(x) = \max_{t \in T^m} h(x,t)$, where T^m is a compact m-dimensional smooth manifold and h: $D \times T^m \to R$ is a q-times continuously differentiable function, $q \geq 2$. Clearly, $S_m^q(D) \subset H(D)$.

Precise information on the local structure of "generic" functions in $S_m^q(D)$ has been obtained by methods of Singularities theory (see e.g. [1], [5], [12], [13]).

However, the following important question seems to be untouched: Do the classes S_m^q really depend on q and m? This question is especially interesting in view of the independence of the class H above of the smoothness of functions involved.

In the present paper we answer this question, showing in many cases noncoincidence of S_m^q for different q and m, although our results are not strong enough to separate these classes completely.

Two different approaches are used: first we study the "massiveness" of the sets S_m^q in appropriate functional spaces, in a way similar to that used in [4], [10] for the problem of representability and approximations by means of compositions. This method allows us to separate classes S_m^q, but does not give explicit examples.

Another approach is based on the study of geometric invariants of maximum functions. These invariants are related to the structure of the set of critical values of the smooth family that defines the maximum function. In this way we obtain explicit examples of functions in $S_m^q \setminus S_{m'}^{q'}$.

To give the flavor of these examples, we state here the following theorem which will be proved in §4 below.

*Received by the editors December 1, 1983, and in revised form October 4, 1984.

† Department of Mathematics, Ben-Gurion University of the Negev, Beer-Sheva 84120, Israel.

THEOREM 4.4. *Let $f(x)$ be a convex piecewise linear function of the single variable $x\in[0,1]$, with the countable number of "edges". Let $\delta_1<\delta_2<\cdots<\delta_i<\cdots$ be the slopes of these edges, and let $\alpha_i=\delta_{i+1}-\delta_i$. Then f can be represented in a form*

$$f(x)=\max_{t\in[0,1]} a(t)x+b(t), \qquad x\in[0,1],$$

with $a(t)$ and $b(t)$ k times continuously differentiable functions on $[0,1]$, if and only if $\sum_{i=1}^{\infty}\alpha_i^{1/k}<\infty$.

There are many open questions, concerning the structure of maximum functions, some of which we discuss in the last section.

2. ε-entropy of sets of maximum functions. Let $D\subset R^n$ be a bounded closed domain. For $q=p+\alpha$, $p\geqq 1$ -an integer, $0<\alpha\leqq 1$, we denote by $C^q(D)$ the space of p times continuously differentiable functions g on D, whose derivatives of order p satisfy the Hölder condition

$$(*) \qquad \|d^p g(x)-d^p g(y)\|\leqq L\|x-y\|^{\alpha},$$

with some constant L.

Let

$$M_i(g)=\max_{y\in D}\|d^i g(y)\|, \qquad i=0,\cdots,p,$$

$$M_q(g)=\inf L \quad \text{in } (*).$$

(We consider all the Euclidean spaces R^s and the spaces of their linear and multilinear mappings with the usual Euclidean norms.)

For $c>0$ denote by $C^q(D,c)$ the set of all g in $C^q(D)$ with $M_i(g)\leqq c$, $i=0,\cdots,p,q$.

Let F be a relatively compact set in a metric space X. For some $\varepsilon>0$ a set $F^*\subset X$ is called an ε-net of F if for any $z\in F$ there exists $z^*\in F^*$ with $d(z,z^*)\leqq\varepsilon$, where d denotes the distance in X.

Denote by $N_\varepsilon(F)$ the number of elements in a minimal ε-net of F. The number $H_\varepsilon(F)=\log_2 N_\varepsilon(F)$ is called the ε-entropy of the set F. It is convenient also to define the number $\mathrm{fd}(F)$ as $\overline{\lim}_{\varepsilon\to 0}\log_2 H_\varepsilon(F)/\log_2(1/\varepsilon)$.

The notion of entropy arises in a natural way in connection with various problems of analysis (see e.g. [4], [10], [14], [16]).

In this section the metric space X is the space $C(D)$ of continuous functions on D with the uniform norm. Now we turn back to maximum functions. Without loss of generality we can assume that the parameter manifold T^m is the unit m-dimensional cube I^m. For $c>0$ denote by $S_m^q(D,c)$ the set of functions f on D, representable as $f(x)=\max_t h(x,t)$, with $h\in C^q(D\times I^m,c)$. Clearly, for any $c>0$, $S_m^q(D,c)$ is a compact subset in $C(D)$.

THEOREM 2.1. *For any $c>0$,*

$$\frac{n+m}{q+2m/n}\leqq \mathrm{fd}\left(S_m^q(D,c)\right)\leqq\frac{n+m}{q}.$$

Proof. Consider the mapping μ: $C(D\times I^m)\to C(D)$, defined by $\mu(h)=f$, $f(x)=\max_{t\in I^m}h(x,t)$. Clearly, μ does not increase the distance. From the definition of ε-entropy we obtain that for any relatively compact set $F\subset C(D\times I^m)$, $H_\varepsilon(\mu(F))\leqq H_\varepsilon(F)$.

Since $S_m^q(D,c)=\mu(C^q(D\times I^m,c))$ and since according to [10, Thm. 2.2.1], $\mathrm{fd}(C^q(D\times I^m,c))=(n+m)/q$, we obtain the right-hand side inequality.

To obtain the lower bound for $\mathrm{fd}(S_m^q(D,c))$ it is sufficient to find in this set a suitable number of functions, any two of which differ at least by 2ε in C-norm.

Let us fix some $\delta>0$ and let $\delta'=\frac{1}{2}\delta^{1+m/n}$. Let $Z_\delta\subset R^n$ denote the net of points of the form $z=(k_1\delta,k_2\delta,\cdots,k_n\delta)$, $k_i\in Z$.

Consider also the points $z'_\alpha\in R^n$ of the form $(k_1\delta',\cdots,k_n\delta')$, $1\leqq k_i\leqq[\delta/\delta']$, indexed in some fixed way, $1\leqq\alpha\leqq[\delta/\delta']^n$. Finitely, let $z''_\beta\in I^m$ be the points of the form $(k_1\delta,\cdots,k_m\delta)$, $0\leqq k_i\leqq[1/\delta]$, indexed in some fixed way, $1\leqq\beta\leqq[1/\delta]^m$.

Since $\delta/\delta'=2(1/\delta)^{m/n}$, $[\delta/\delta']^n\sim 2^n(1/\delta)^m>[1/\delta]^m$ for δ sufficiently small, we can fix some one-to-one mapping ω from the set of indices β into the set of indices α.

Now consider in $R^m\times I^m$ the net of points y of the form $y=(z+z'_{\omega(\beta)},z''_\beta)$, $z\in Z_\delta$, $1\leqq\beta\leqq[1/\delta]^m$, and let y_ξ, $1\leqq\xi\leqq K(\delta)$, denote those from the points y, whose projections x_ξ on R^n belong to D. Since D has a nonempty interior, $K(\delta)\geqq K'(1/\delta)^{n+m}$, with some $K'>0$, not depending on δ.

We shall use below only the following property of points y_ξ: they form a net in $D\times I^m$ with the distance at least δ between any two points, while their projections x_ξ form a net in D with the distance at least δ' between any two points.

Let ϕ: $R\to R$ be a C^∞-smooth even function with the following properties:

i. $\phi(s)=1-s^2$ for $|s|\leqq\frac{1}{2}$;

ii. $\phi(s)=0$ for $|s|\geqq 1$;

iii. $\phi(s)$ is a decreasing function of $|s|$ for $\frac{1}{2}\leqq|s|\leqq 1$.

Let ψ: $R^{n+m}\to R$ be defined by $\psi(y)=\phi(\|y\|)$. Denote by M $2^q\cdot\max M_i(\psi)$, $i=1,\cdots,p+1$.

Now for any subset $\kappa\subset\{1,\cdots,K(\delta)\}$ define the function ψ_κ: $D\times I^m\to R$ as follows: for $y=(x,t)\in D\times I^m$

$$\psi_\kappa(y)=\frac{c}{M}\delta^q\sum_{\xi\in\kappa}\psi\left(\frac{2}{\delta}(y-y_\xi)\right).$$

Clearly, ψ_κ is a C^∞ function, and since the supports of $\psi((2/\delta)(y-y_\xi))$ are disjoint for different ξ by property ii of ϕ, we see that $\psi_\kappa\in C^q(D\times I^m,c)$.

Now consider the corresponding maximum function $f_\kappa=\mu(\psi_\kappa)$, which by definition belongs to $S_m^q(D,c)$.

LEMMA 2.2. *For* $\kappa\neq\kappa'$, $\|f_\kappa-f_{\kappa'}\|_C\geqq(c/M)\delta^{q+2m/n}$.

Proof. Since the supports of $\psi((2/\delta)(y-y_\xi))$ are disjoint, we have:

$$f_\kappa(x)=\frac{c}{M}\delta^q\max_{\xi\in\kappa}\max_{t\in I^m}\psi\left(\frac{2}{\delta}(y-y_\xi)\right)=\frac{c}{M}\delta^q\max_{\xi\in\kappa}\phi\left(\frac{2}{\delta}\|x-x_\xi\|\right),$$

by construction of function ψ.

Now assume that $\kappa\setminus\kappa'\neq\phi$ and fix some $\eta\in\kappa\setminus\kappa'$. Then

$$f_\kappa(x_\eta)=\frac{c}{M}\delta^q\phi(0)=\frac{c}{M}\delta^q,$$

by property i of ϕ. On the other hand, for any $\xi \neq \eta$,

$$\phi\left(\frac{2}{\delta}\|x_\eta - x_\xi\|\right) = 1 - \left(\frac{2}{\delta}\|x_\eta - x_\xi\|\right)^2 \leqq 1 - 4\left(\frac{\delta'}{\delta}\right)^2,$$

since $\|x_\eta - x_\xi\| \geqq \delta'$ for $\xi \neq \eta$. Hence

$$f_{\kappa'}(x_\eta) \leqq \frac{c}{M}\delta^q - \frac{c}{M}\delta^q 4\left(\frac{\delta'}{\delta}\right)^2 = f_\kappa(x_\eta) - \frac{c}{M}\delta^{q+2m/n}.$$

Therefore $\|f_\kappa - f_{\kappa'}\|_C \geqq \|f_\kappa(x_\eta) - f_{\kappa'}(x_\eta)\| \geqq (c/M)\delta^{q+2m/n}$. This completes the proof of Lemma 2.2. We now return to the proof of Theorem 2.1.

Given $\varepsilon > 0$ let $\delta = (3M\varepsilon/c)^{1/(q+2m/n)}$. Then $(c/M)\delta^{q+2m/n} = 3\varepsilon$ and by Lemma 2.2 all the functions f_κ form the 3ε-separated net in $S_m^q(D,c)$. Since the number of elements in this net is equal to

$$2^{K(\delta)} \geqq 2^{K'(1/\delta)^{n+m}} = 2^{K''(1/\varepsilon)^{(n+m)/(q+2m/n)}},$$

we have

$$H_\varepsilon(S_m^q(D,c)) \geqq K''(1/\varepsilon)^{(n+m)/(q+2m/n)}.$$

Theorem 2.1 is proved.

As an immediate consequence of Theorem 2.1 we obtain the main result of this section:

THEOREM 2.3. *Let D be a compact domain in R^n. Then for any m, q and m', q', such that $(n+m)/(q+2m/n) > (n+m')/q'$, the set of functions in $S_m^q(D) \subset C(D)$ not belonging to $S_{m'}^{q'}(D)$, is a set of second category. In particular, this set is everywhere dense in $S_m^q(D)$.*

Proof. Clearly, $S_{m'}^{q'}(D) = \bigcup_{N=1}^{\infty} S_{m'}^{q'}(D,N)$. By Theorem 2.1 for any ball Ω in $C(D)$, $\mathrm{fd}(S_m^q(D) \cap \Omega) \geqq (n+m)/(q+2m/n)$. But by the same theorem, $\mathrm{fd}(S_{m'}^{q'}(D,N)) \leqq (n+m')/q' < (n+m)/(q+2m/n)$. Since $S_{m'}^{q'}(D,N) \cap S_m^q(D)$ is closed, it is therefore a nowhere dense subset in S_m^q.

Now we give some corollaries, showing what classes S_m^q can be separated by means of Theorem 2.3.

COROLLARY 2.4. *For given $D \subset R^n$ and m and for any q and $q' > q + 2m/n$, $S_m^{q'}(D) \subsetneqq S_m^q(D)$. In particular, if $n > 2m$, then for any $q' > q$, $S_m^{q'}(D) \subsetneqq S_m^q(D)$.*

Proof. For $q' > q + 2m/n$, $(n+m)/q' < (n+m)/(q+2m/n)$.

COROLLARY 2.5. *For given $D \subset R^n$, q and m, and for any $m' < m - 2n(n+m)/(2n+qm)$, $S_{m'}^q(D) \subsetneqq S_m^q(D)$. In particular, for $q > (2n^2-1)/m + 2n$, $S_{m'}^q(D) \subsetneqq S_m^q(D)$ for any $m' < m$, and for $q > 2n^2 + 2n - 1$, $S_{m'}^q(D) \subsetneqq S_{m''}^q(D)$ for any $m' < m''$.*

Proof. For $m' < m - 2n(n+m)/(2n+qm)$, $(n+m')/q < (n+m)/(q+2m/n)$. For $q > (2n^2-1)/m + 2n$, $2n(n+m)/(2n+qm) < 1$. Finitely, for $q > 2n^2 + 2n - 1$, the last inequality is satisfied for any m.

We can use the result of Theorem 2.1 also to compare $S_m^q(D)$ with $C^k(D)$. Indeed, by [10, Thm. 2.2.1], $\mathrm{fd}\, C^k(D) = n/k$, and repeating the proof of Theorem 2.3 we obtain the following:

COROLLARY 2.6. *For $(n+m)/q < n/k$, the set of k times continuously differentiable functions on D not belonging to $S_m^q(D)$, is a set of the second category and, in particular, is everywhere dense in the uniform topology.*

Notice that Corollaries 2.4, 2.5 and 2.6 do not give explicit examples of functions, not representable as a maximum of a suitably smooth family. Below we give such examples for any situation, covered by Corollary 2.6.

3. Critical values of maximum functions. Let $f\colon D \to R^s$ be a continuous mapping. The point $x \in \mathring{D}$ is called a critical point of f, if the first differential $df(x)$ exists and is equal to zero. Let $\Sigma(f)$ be the set of all the critical points of f and let $\Delta(f)=f(\Sigma(f)) \subset R^s$ be the set of critical values of f. (Actually, $\Sigma(f)$ is the set of critical points of rank zero of f, in the usual terminology).

A well-known and widely used property of critical values of differentiable mappings is given by the Morse–Sard theorem (see [3], [8]): if the mapping is sufficiently smooth, the set of its critical values has the Lebesgue measure (or, more precisely, the Hausdorff measure of an appropriate dimension) zero.

In [14] the stronger property of critical values has been established: let A be a bounded subset in R^s. The entropy dimension of A, $\dim_e A$ is defined as

$$\dim_e A = \inf\left\{ \beta\colon \exists K, \forall \varepsilon > 0,\ \varepsilon \leqq 1,\ N_\varepsilon(A) \leqq K\left(\frac{1}{\varepsilon}\right)^\beta \right\},$$

where $N_\varepsilon(A)$, as above, is the number of elements in a minimal ε-net of A in R^s.

For "nice" sets the entropy dimension coincides with the Hausdorff dimension $\dim_h$ and with the topological dimension. For any A $\dim_e A \geqq \dim_h A$, and the important advantage of the entropy dimension, which we shall use below, is that it allows us to distinguish countable sets, while the Hausdorff dimension of any countable set is zero.

The following result has been obtained in [14]:

THEOREM 5.4. [14]. *Let $D \subset R^n$ be a compact domain and let $f\colon D \to R^s$ be a C^q-mapping. Then*

$$\dim_e \Delta(f) \leqq \frac{n}{q}.$$

(The Morse–Sard theorem, in its general form, proved in [3], gives the same bound for the Hausdorff dimension of $\Delta(f)$.)

It turns out that also the critical values of the function, representable as a maximum of a smooth family, cannot form a set which is too large. More precisely, we have:

THEOREM 3.1. *Let $D \subset R^n$ be a compact domain. For any $f \in S_m^q(D)$,*

$$\dim_e \Delta(f) \leqq \frac{n+m}{q}.$$

Proof. Let $f(x)=\max_{t \in I^m} h(x,t)$, $h \in C^q(D \times I^m)$.

Before proving Theorem 3.1 we state and prove a lemma.

LEMMA 3.2. *Let $x_0 \in D$ be a critical point of f and let $t_0 \in I^m$ be such that $f(x_0)=h(x_0,t_0)$. Then $(x_0,t_0) \in D \times I^m$ is a critical point of h.*

Proof. Since $h(x_0,t)$ attains its maximum with respect to t at t_0, we have $d_t h(x_0,t_0)=0$.

By the definition of critical points, f is differentiable at x_0 and $df(x_0)=0$. But then from the expression of the generalized differential of maximum functions (see [2]) it follows immediately that $d_x h(x_0,t_0)=0$.

Proof of Theorem 3.1. By Lemma 3.2 any critical value $f(x_0)$ of f is a critical value $h(x_0,t_0)$ of h, i.e., $\Delta(f) \subset \Delta(h)$. But by Theorem 5.4 [14], $\dim_e \Delta(h) \leqq (n+m)/q$.

Now to obtain examples of functions, nonrepresentable as maximum, we note that for any $\gamma < n/k$, by [14, Thm. 5.6], the function $g \in C^k(D)$ can be built, with $\dim_e \Delta(g) \geqq \gamma$. By Theorem 3.1, g does not belong to any $S_m^q(D)$, with $(n+m)/q < \gamma$; and therefore, we get explicit examples of nonrepresentable functions in any situation, covered by Corollary 2.6.

In particular, let g_n: $I^n \to R$, $g_n \in C^{n-1}(I^n)$ be the Whitney function with $\Delta(g) = [0,1]$. (See [11].)

Since $\dim_e [0,1] = 1$, we have proved:

COROLLARY 3.3. *$g_n \notin S_m^q(I^n)$ for $q > n+m$.*

Notice that in all the constructions above it is sufficient to use the Hausdorff dimension of the sets of critical values. However, using the specific properties of the entropy dimension we can give examples of very simple functions, not representable as maximum:

Let ψ_k: $[0,1] \to R$ be defined as $\psi_k(x) = x^k \cos(1/x)$. $\psi_k \in C^{[k/2]-1}([0,1])$ and the critical values of ψ_k form the sequence

$$\omega\left(-\left(\frac{1}{\pi}\right)^k\right), \omega\left(\left(\frac{1}{2\pi}\right)^k\right), \cdots, \omega\left((-1)^i\left(\frac{1}{i\pi}\right)^k\right), \cdots,$$

where ω: $[-1,1] \to [-1,1]$ is a Lipschitzian homeomorphism. Hence $\dim_e \Delta(\psi_k) = 1/(k+1)$ (see e.g. [15]). This proves:

COROLLARY 3.4. *$\psi_k \notin S_m^q([0,1])$ for $q > (k+1)(m+1)$.*

4. Maxima of smooth families of linear functions. In this section we give simple examples of convex functions nonrepresentable as a maximum of a sufficiently smooth family of linear functions. Once more we reduce the question of representability to the properties of critical values of some differentiable mappings. But here, in constrast to section 3, the arising sets of critical values are a priori at most countable. Thus the Morse–Sard theorem gives no information in this case and the use of the entropy dimension and the stronger [14, Thm. 5.4] becomes essential. In fact, this theorem was found in the attempt to give criteria for representability of convex functions as the maximum of linear ones.

Let D be a convex compact domain in R^n. We consider a cone $Q(D)$ of convex functions f on D, which are extendable to convex functions on R^n and whose graphs $\Gamma(f)$ over D are polyhedra with possibly countable number of faces.

We study the representability of $f \in Q(D)$ as $f = \max_{t \in I^m} l_t$, where for $x = (x_1, \cdots, x_n) \in R^n$

$$l_t(x) = a_1(t)x_1 + \cdots + a_n(t)x_n + b(t), \quad \text{with } a_1, \cdots, a_n, b \in C^q(I^m).$$

Denote the set of functions in $Q(D)$, representable in this form, by $Q_m^q(D)$.

Let L be a hyperplane in $R^n \times R$ with a nondegenerate projection on R^n. L is a graph of the linear function $l_L(x) = a_1(L)x_1 + \cdots + a_n(L)x_n + b(L)$. Denote the point $(a_1(L), \cdots, a_n(L)) \in R^n$ by $\delta(L)$.

For a given $f \in Q(D)$ let $\delta(F) \subset R^n$ be the set of $\delta(\gamma)$, where γ runs through all the faces of $\Gamma(f)$. $\delta(f)$ is at most countable bounded subset in R^n.

THEOREM 4.1. *For any $f \in Q_m^q(D)$, $\dim_e \delta(F) \leqq (m/q)$.*

Proof. Write f as $\max l_t$, $l_t(x) = a_1(t)x_1 + \cdots + a_n(t)x_n + b(t)$, and for each face γ of $\Gamma(f)$ find some $t_\gamma \in I^m$, such that γ is a graph of l_{t_γ}.

LEMMA 4.2. *For any face γ of $\Gamma(f)$, t_γ is a critical point of each of functions $a_1, \cdots, a_n, b$.*

Proof. Let $x^0, x^1, \cdots, x^n$ be the vertices of some nondegenerate simplex in the projection of γ on R^n. Since at each x^i, $f(x^i)=l_{t_\gamma}(x^i)=\max_t l_t(x^i)$, we have $d_t l_t(x^i)\,|_{t=t_\gamma}=0$, or

$$d_t a_1(t_\gamma)x_1^i+\cdots+d_t a_n(t_\gamma)x_n^i+d_t b(t_\gamma)=0, \qquad i=0,\cdots,n.$$

But since $x^0,\cdots,x^n$ are the vertices of a nondegenerated simplex, this linear system has only the zero solution. This proves the lemma.

Thus if we define the mapping Φ: $I^m\to R^n$ by $\Phi(t)=(a_1(t),\cdots,a_n(t))$, any point t_γ is a critical point of Φ and $\delta(\gamma)=\Phi(t_\gamma)$ is a critical value of Φ. Hence $\delta(f)\subset\Delta(\Phi)$, and since by Theorem 5.4 [14] $\dim_e\Delta(\Phi)\leqq m/q$, Theorem 4.1 is proved.

Now consider, for instance, the set $\Delta_\alpha\subset R^n$, consisting of points of the form

$$\left(\frac{1}{k_1^\alpha},\frac{1}{k_2^\alpha},\cdots,\frac{1}{k_n^\alpha}\right), \qquad k_i=1,2,\cdots,$$

$\alpha>0$. We have $\dim_e\Delta_\alpha=n/(\alpha+1)$ (see e.g. [15]). One can easily find functions $f_\alpha\in Q(I^n)$ with $\delta(f)=\Delta_\alpha$. We obtain the following:

COROLLARY 4.3. *$f_\alpha\notin Q_m^q(I^n)$ for $m/q<n/(\alpha+1)$.*

Using the metric invariant V_β, defined in [15], one can give a more precise version of Theorem 4.1. We shall consider only one special case, where the question of representability can be answered completely. This is the case of Theorem 4.4 which was stated in the introduction. We repeat that statement here.

Let $f\in Q([0,1])$ be a function, for which $\delta(f)$ is a sequence $\delta_1<\delta_2<\cdots<\delta_i<\cdots$ and let $\alpha_i=\delta_{i+1}-\delta_i$.

THEOREM 4.4. *The function f can be represented as $f(x)=\max_{t\in[0,1]}a(t)x+b(t)$, $x\in[0,1]$, with a and b k times continuously differentiable functions on $[0,1]$, if and only if $\Sigma_{i=1}^\infty\alpha_i^{1/k}<\infty$.*

Proof. If f has a required representation, then, by Lemma 4.2, $\delta(f)=\{\delta_1,\delta_2,\cdots\}\subset\Delta(a)$. But then by [15, Thm. 4.1], $\Sigma_{i=1}^\infty\alpha_i^{1/k}<\infty$.

Now assume that $\Sigma_{i=1}^\infty\alpha_i^{1/k}<\infty$. In the proof of [15, Thm. 4.1] it is shown that we can find the function a: $[0,1]\to R$ with the following properties:

(i) $a\in C^\infty([0,1))$ and all the derivatives of a up to order k at $t\in[0,1)$ tend to zero as t tends to 1 (and, in particular, a is k times continuously differentiable on $[0,1]$.)

(ii) a increases on $[0,1]$.

(iii) There is a sequence of points $t_1, t_2,\cdots$ in $[0,1)$, converging to 1, such that $a(t_i)=\delta_i$ and all the derivatives of a at t_i vanish, $i=1,2,\cdots$.

Using $a(t)$ and f, we define $b(t)$ as follows: $b(t)$ is the constant term in the equation of the support line to the graph $\Gamma(f)$ with the slope $a(t)$.

Clearly, $f(x)=\max_t a(t)x+b(t)$. It remains to prove that b is k times continuously differentiable on $[0,1]$.

Let (x_i,y_i) be the coordinates of the vertex of $\Gamma(f)$, belonging to the edges of $\Gamma(f)$ with the slopes δ_i and δ_{i+1}, $i=1,2,\cdots$. Then for $\delta_i\leqq a(t)\leqq\delta_{i+1}$, i.e. for $t_i\leqq t\leqq t_{i+1}$, $b(t)=y_i-a(t)x_i$. Hence $b(t)$ is C^∞-smooth on each segment $[t_i,t_{i+1}]$, and its derivatives coincide with the derivatives of $a(t)$ up to a coefficient $-x_i$. But by the condition (iii), all the derivatives of a vanish at t_i. Hence $b\in C^\infty([0,1])$. Since by i, all the derivatives of $a(t)$ up to order k tend to zero as t tends to 1, the same is true for $b(t)$ and hence b is a k times continuously differentiable function on $[0,1]$.

5. Some open questions. Of course, the results above are far from being complete. However, they show that there is a rich variety of interesting phenomena concerning the maxima of smooth families, as well as various connections with the deep properties of smooth mappings.

Theorem 2.1 gives bounds for the functional dimension fd of the sets of maximum functions. What is the precise value of $\mathrm{fd}(S_m^q(D,c))$?

Notice also that the invariants obtained in §§2 and 3 involve both the smoothness q and the dimension of the parameter space m. Can one find invariants of maximum functions, separating the influence of these factors?

There is a similarity between the study of maximum functions above and the study of functions, representable by means of compositions of functions of some fixed class (see [10], [16]). In both cases the consideration of the ε-entropy allows us to prove the existence of nonrepresentable functions, while the study of some invariants, related to critical values, gives explicit examples of such functions. Are there direct connections between these two problems?

The necessary condition for the representability of a polyhedral convex function as a maximum of a smooth family of linear functions, given by Theorem 4.1, is very close to a sufficient one (see Theorem 4.4). However, for an arbitary convex function (not necessarily polyhedral), the method used here breaks down. Can one define invariants of a general convex function, responsible for the representability of this function as a maximum of a smooth family of linear functions?[1]

The maximum functions of smooth families have nice differentiability properties, which can be formulated, in particular, in terms of their generalized derivatives (both in sense of distributions—see [12], and in sense of optimization theory—see [2], [6], [7], [9]). Can one give criteria of representability in these terms? In particular, can one find function spaces appropriate for the treatment of maximum functions?

We focus on one question, in particular concerning the differentiability properties of maximum functions. A continuous function $f\colon R^n \to R$ is said to have the kth differential at x_0, if there exists a polynomial $P\colon R^n \to R$ of degree k, such that $\|f(x)-P(x)\| = o(\|x-x_0\|^k)$. Convex functions are known to have the second differential almost everywhere. Is it true that functions in $S_m^k(D)$ have the kth differential almost everywhere in D? Some variant of this question (for finite families) is considered in [17].[2]

Acknowledgment. The author would like to thank Y. Kanai for useful discussions and the Max-Planck-Institut für Mathematik, where this paper was written, for its kind hospitality.

REFERENCES

[1] L. N. BRYZGALOVA, *Singularities of a maximum of a function depending on parameters*, Funct. Anal. Appl., 11 (1977), pp. 49–50.

[2] F. H. CLARKE, *Generalized gradients and applications*, Trans. Amer. Math. Soc., 205 (1975), 247–262.

[3] H. FEDERER, *Geometric Measure Theory*, Die Grundlehren der Mathematischen Wissenschaften, Springer, Berlin, Heidelberg, New York, 1969.

[1] Some necessary conditions for representability of general convex functions are obtained in: Y. Yomdin, *On representability of convex functions as maxima of linear families*, to appear.

[2] The answer to this question is positive, at least in the sense that functions in $S_m^k(D)$ have the $l(k)$th differential almost everywhere in D, with $l(k)\to\infty$ as $k\to\infty$.

[4] A. N. KOLMOGOROV AND V. M. TIKHOMIROV, *ε-entropy and ε-capacity of sets in functional spaces*, Uspekhi Mat. Nauk, 14, 2 (1959), pp. 3–86; Amer. Math. Soc. Transl., 72 (1961), pp. 277–364.

[5] V. I. MATOV, *Topological classification of germs of maximum and minimax functions of generic families*, Uspekhi Mat. Nauk 37, 4 (1982), pp. 167–168. (In Russian)

[6] B. N. PSCHENICHNYI, *Necessary conditions for an extremum*, Marcel Dekker, New York, 1971.

[7] R. T. ROCKAFELLAR, *Favorable classes of Lipschitz continuous functions in subgradient optimization*, in Progress in Nondifferentiable Optimization, E. Nurminski, ed., Pergamon Press, New York, 1981.

[8] A. SARD, *The measure of the critical values of differential maps*, Bull. Amer. Math. Soc., 48 (1942), 883–890.

[9] A. SHAPIRO AND Y. YOMDIN, *On functions representable as a difference of two convex functions, and necessary conditions in a constrained optimization*, preprint, 1981.

[10] A. G. VITUSHKIN, *On representation of functions by means of superpositions and related topics*, Enseignement Math., 23 (1977), pp. 255–320.

[11] H. WHITNEY, *A function not constant on a connected set of critical points*, Duke Math. J., 1 (1935), pp. 514–517.

[12] Y. YOMDIN, *On functions representable as a supremum of a family of smooth functions*, this Journal, 14 (1983), pp. 239–246.

[13] _____, *On the local structure of a generic central set*, Compositio Math., 43 (1981), pp. 225–238.

[14] _____, *The geometry of critical and near-critical values of differentiable mappings*, Math. Ann., 264 (1983), pp. 495–515.

[15] _____, *β-spread of sets in metric spaces and critical values of smooth functions*, to appear.

[16] _____, *Critical values and representation of functions by means of compositions*, this Journal, 17 (1986), pp. 236–239.

[17] _____, *Some results on finite determinancy and stability not requiring the explicit use of smoothness*, Proc. Symposia on Pure Mathematics, 40 (1983), pp. 667–674.

SIAM J. MATH. ANAL.
Vol. 17, No. 4, July 1986

015

POSITIVITY OF THE POISSON KERNEL FOR THE CONTINUOUS q-JACOBI POLYNOMIALS AND SOME QUADRATIC TRANSFORMATION FORMULAS FOR BASIC HYPERGEOMETRIC SERIES*

GEORGE GASPER† AND MIZAN RAHMAN‡

Abstract. A nonterminating extension of the Sears–Carlitz quadratic transformation formula for a well-posed ${}_3\phi_2$ series with an arbitrary argument is obtained as a sum of two balanced ${}_5\phi_4$ series. This is then extended to a very well-poised ${}_5\phi_4$ series with arbitrary argument. These results are used to derive some generating functions for the q-Wilson polynomials $p_n(x;\ a,b,c,d;\ q)$ when $ad=bc$ and an expression for the Poisson kernel $K_t(x,y;\ a,b,c,bc/a;\ q)$ as a sum of three sums of very well-poised ${}_{10}\phi_9$ series which clearly demonstrates its positivity for $0\leqq t<1$, $0\leqq q<1$ in the continuous q-Jacobi case when $a=q^{\alpha/2+1/4}$, $b=q^{\alpha/2+3/4}$, $c=-q^{\beta/2+1/4}$ and $\alpha,\beta>-1$. Additional quadratic transformation formulas are derived, along with q-analogues of Watson's and Whipple's summation formulas.

Key words. q-Wilson polynomials, continuous q-Jacobi polynomials, Sears–Carlitz formula, quadratic transformations of basic hypergeometric series, q-analogues of Watson's and Whipple's summation theorems, Bailey's transformation formulas, generating functions, Poisson kernel

AMS(MOS) subject classifications. Primary 33A65, 33A70; secondary 33A30

1. Introduction.

Recently we showed [9] how Rogers' linearization formula for the continuous q-ultraspherical polynomials $C_n(x;\ \beta\,|\,q)$, defined below, could be used to derive an ${}_8\phi_7$ basic hypergeometric series representation for the Poisson kernel $P_t(x,y;\ \beta\,|\,q)$ for these polynomials from which one could easily see that this kernel is positive for $-1\leqq x,y\leqq 1$, $0\leqq t,q<1$ when $0\leqq\beta<1$. In analogy with the limiting ultraspherical polynomial case we conjectured that $P_t(x,y;\ \beta\,|\,q)$ should also be positive when $1<\beta<q^{-1/2}$. It was in trying to prove this conjecture and to give conditions under which the Poisson kernel for the continuous q-Jacobi polynomials $P_n^{(\alpha,\beta)}(x\,|\,q)$ is positive that we were led to consider the quadratic transformations and generating functions derived in this paper.

In particular, after first introducing our notation in §2, we shall prove in §3, that the Sears–Carlitz formula [14, (4.1)], [8, (2.4)],

$$(1.1)\quad {}_3\phi_2\left[\begin{matrix}a,\ b,\ c\\ aq/b,\ aq/c\end{matrix};q,\ \frac{aqx}{bc}\right]$$

$$=\frac{(ax;\ q)_\infty}{(x;\ q)_\infty}\ {}_5\phi_4\left[\begin{matrix}\sqrt{a},\ -\sqrt{a},\ \sqrt{aq},\ -\sqrt{aq},\ aq/bc\\ aq/b,\ aq/c,\ ax,\ q/x\end{matrix};q,\ q\right]$$

*Received by the editors November 8, 1983, and in revised form August 29, 1985.

†Department of Mathematics, Northwestern University, Evanston, Illinois 60201. The work of this author was supported in part by the National Science Foundation under grant MCS-8002507 A01.

‡Department of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada K1S 5B6. The work of this author was supported in part by the Natural Sciences and Engineering Research Council under grant A6197.

where $a=q^{-n}$, $n=0,1,\cdots$, has a nonterminating extension of the form

$$(1.2)\quad {}_3\phi_2\left[\begin{matrix}a, b, c\\ aq/b, aq/c\end{matrix};q, \frac{aqx}{bc}\right]$$

$$=\frac{(ax;q)_\infty}{(x;q)_\infty}\,{}_5\phi_4\left[\begin{matrix}\sqrt{a}, -\sqrt{a}, \sqrt{aq}, -\sqrt{aq}, aq/bc\\ aq/b, aq/c, ax, q/x\end{matrix};q, q\right]$$

$$+\frac{(a;q)_\infty(aq/bc;q)_\infty(aqx/b;q)_\infty(aqx/c;q)_\infty}{(aq/b;q)_\infty(aq/c;q)_\infty(aqx/bc;q)_\infty(x^{-1};q)_\infty}$$

$$\cdot\,{}_5\phi_4\left[\begin{matrix}x\sqrt{a}, -x\sqrt{a}, x\sqrt{aq}, -x\sqrt{aq}, aqx/bc\\ aqx/b, aqx/c, xq, ax^2\end{matrix};q, q\right].$$

An analogous quadratic transformation formula, (3.5) below, is derived for the series

$$(1.3)\quad {}_5\phi_4\left[\begin{matrix}a, q\sqrt{a}, -q\sqrt{a}, b, c\\ \sqrt{a}, -\sqrt{a}, aq/b, aq/c\end{matrix};q, \frac{t\sqrt{aq}}{bc}\right],$$

which plays a crucial role in our derivation in §6 of a representation for the Poisson kernel for the continuous q-Jacobi polynomials as a sum of three sums of balanced ${}_{10}\phi_9$ series from which its positivity for $0\leq t,q<1$ is obvious when $\alpha,\beta>-1$; thus, also proving the above-mentioned conjecture for the continuous q-ultraspherical polynomials.

We shall also derive quadratic transformations of the type that transform series with base q^2 to series with base q, such as

(1.4)

$${}_2\phi_1\left[\begin{matrix}a^2, b^2\\ a^2q^2/b^2\end{matrix};q^2, \frac{x^2q^2}{b^4}\right]=\frac{(aq/b^2;q)_\infty(x^2q/b^2;q)_\infty}{(q/b^2;q)_\infty(ax^2q/b^2;q)_\infty}$$

$$\cdot\frac{(a^2x^2q/b^2;q^2)_\infty(q/b^2;q^2)_\infty(q^2/b^2;q^2)_\infty(a^2x^2q^2/b^4;q^2)_\infty}{(x^2q/b^2;q^2)_\infty(a^2q/b^2;q^2)_\infty(a^2q^2/b^2;q^2)_\infty(x^2q^2/b^4;q^2)_\infty}$$

$$\cdot\,{}_8\phi_7\left[\begin{matrix}ax^2/b^2;q\sqrt{\cdot}, -q\sqrt{\cdot}, a, x, -x, xq^{1/2}/b, -xq^{1/2}/b\\ \sqrt{\cdot}, -\sqrt{\cdot}, x^2q/b^2, axq/b^2, -axq/b^2, axq^{1/2}/b, -axq^{1/2}b\end{matrix};q, \frac{aq}{b^2}\right],$$

where $|aq/b^2|<1$, and give q-analogues of the well-known quadratic summation formulas of Watson [5, p. 16].

$$(1.5)\quad {}_3F_2\left[\begin{matrix}a, b, c\\ \frac{1}{2}(a+b+1), 2c\end{matrix};1\right]=\frac{\Gamma(\frac{1}{2})\Gamma(\frac{1}{2}+c)\Gamma(\frac{1+a+b}{2})\Gamma(\frac{1-a-b+c}{2})}{\Gamma(\frac{1+a}{2})\Gamma(\frac{1+b}{2})\Gamma(\frac{1-a}{2}+c)\Gamma(\frac{1-b}{2}+c)}$$

and Whipple [5, p. 16]

$$(1.6)\quad {}_3F_2\left[\begin{matrix}a, 1-a, b\\ c, 2b+1-c\end{matrix};1\right]=\frac{\pi\Gamma(c)\Gamma(2b+1-c)2^{1-2b}}{\Gamma(\frac{a+c}{2})\Gamma(b+\frac{a+1-c}{2})\Gamma(\frac{1+c-a}{2})\Gamma(b+1-\frac{a+c}{2})}.$$

The open square root symbol used in (1.4) and elsewhere is an abbreviation for the square root of the top left-hand member of the series, with the understanding that the same square root is used throughout.

2. Notation. Let $|q|<1$. As usual, a ${}_{r+1}\phi_r$ basic hypergeometric series with base q is defined by

$$(2.1)\qquad {}_{r+1}\phi_r\left[\begin{matrix}a_1,a_2,\cdots,a_{r+1}\\ b_1,\cdots,b_r\end{matrix};q,\ z\right]=\sum_{n=0}^{\infty}\frac{(a_1;\ q)_n(a_2;\ q)_n\cdots(a_{r+1};\ q)_n}{(q;\ q)_n(b_1;\ q)_n\cdots(b_r;\ q)_n}z^n,$$

whenever the series converges (e.g., if $|z|<1$) where the q-shifted factorials are defined by

$$(a;\ q)_n=\begin{cases}1, & n=0,\\ (1-a)(1-aq)\cdots(1-aq^{n-1}), & n=1,2,\cdots.\end{cases}$$

We also let $(a;\ q)_\infty=(1-a)(1-aq)(1-aq^2)\ldots$. To simplify the notation we will also write $(a)_n$ and $(a)_\infty$ in place of $(a;\ q)_n$ and $(a;\ q)_\infty$, respectively, and omit the base q symbol in the ${}_{r+1}\phi_r$ series when the base is q, so that only the z term appears after the semicolon.

As in Askey and Wilson [4] we define the continuous q-ultraspherical polynomials $C_n(x;\ \beta\,|\,q)$ by

$$(2.2)\qquad C_n(x;\ \beta\,|\,q)=\frac{(\beta^2)_n\beta^{-n/2}}{(q)_n}\,{}_4\phi_3\left[\begin{matrix}q^{-n},\ q^n\beta^2,\ \beta^{1/2}e^{i\theta},\ \beta^{1/2}e^{-i\theta}\\ \beta q^{1/2},\ -\beta q^{1/2},\ -\beta\end{matrix};q\right]$$

and the continuous q-Jacobi polynomials $P_n^{(\alpha,\beta)}(x\,|\,q)$ by

$$(2.3)\qquad P_n^{(\alpha,\beta)}(x\,|\,q)=\frac{(q^{\alpha+1})_n}{(q)_n}\,{}_4\phi_3\left[\begin{matrix}q^{-n},\ q^{n+\alpha+\beta+1},\ q^{\alpha/2+1/4}e^{i\theta},\ q^{\alpha/2+1/4}e^{-i\theta}\\ q^{\alpha+1},\ -q^{\alpha/2+\beta/2+1/2},\ -q^{\alpha/2+\beta/2+1}\end{matrix};q\right],$$

where, as elsewhere, $x=\cos\theta$. It should be noted that $C_n(x;\ q^{\alpha+1/2}\,|\,q)$ is a constant multiple of $P_n^{(\alpha,\alpha)}(x\,|\,q)$ and that the polynomial

$$(2.4)\qquad P_n^{(\alpha,\beta)}(x;\ q)=\frac{(q^{\alpha+1})_n(-q^{\beta+1})_n}{(q)_n(-q)_n}\,{}_4\phi_3\left[\begin{matrix}q^{-n},\ q^{n+\alpha+\beta+1},\ q^{1/2}e^{i\theta},\ q^{1/2}e^{-i\theta}\\ q^{\alpha+1},\ -q^{\beta+1},\ -q\end{matrix};q\right],$$

which was considered in Rahman [13] and Gasper [11], is a constant multiple of $P_n^{(\alpha,\beta)}(x\,|\,q^2)$, as Askey and Wilson showed in [4, §4].

Since the above polynomials are constant multiples of the q-Wilson polynomials [4]

$$(2.5)\qquad p_n(x;\ a,b,c,d;\ q)={}_4\phi_3\left[\begin{matrix}q^{-n},\ abcdq^{n-1},\ ae^{i\theta},\ ae^{-i\theta}\\ ab,\ ac,\ ad\end{matrix};q\right],$$

we shall initially consider these polynomials. Askey and Wilson [4, Thm. 2.2] proved for $-1<q<1$ that if a, b, c, d are real (or if complex, appear in conjugate pairs) and

$\max(|a|, |b|, |c|, |d|)<1$ then these polynomials satisfy the orthogonality relation

$$\int_{-1}^{1} p_n(x; a,b,c,d; q)p_m(x; a,b,c,d; q)w(x; a,b,c,d; q)\,dx=\frac{\delta_{m,n}}{h_n}, \tag{2.6}$$

where

$$w(x; a,b,c,d; q) = \frac{(e^{2i\theta})_\infty(e^{-2i\theta})_\infty \csc\theta}{(ae^{i\theta})_\infty(ae^{-i\theta})_\infty(be^{i\theta})_\infty(be^{-i\theta})_\infty(ce^{i\theta})_\infty(ce^{-i\theta})_\infty(de^{i\theta})_\infty(de^{-i\theta})_\infty}, \tag{2.7}$$

and

$$\begin{aligned} h_n &= h_n(a,b,c,d; q) \\ &= \frac{(abcdq^{-1})_n(1-abcdq^{2n-1})(ab)_n(ac)_n(ad)_n a^{-2n}}{(1-abcdq^{-1})(bc)_n(bd)_n(cd)_n(q)_n} \\ &\quad \cdot \frac{(ab)_\infty(ac)_\infty(ad)_\infty(bc)_\infty(bd)_\infty(cd)_\infty(q)_\infty}{2\pi(abcd)_\infty}. \end{aligned} \tag{2.8}$$

They also gave the orthogonality relation for cases when $a>1$, $0<q<1$, and some positive discrete masses have to be added to the weight function (see [4, Thm. (2.4)]).

From (2.8) the Poisson kernel

$$\begin{aligned} &P_t(x,y; a,b,c,d; q) \\ &\quad = \sum_{n=0}^{\infty} t^n h_n(a,b,c,d; q)p_n(x; a,b,c,d; q)p_n(y; a,b,c,d; q) \end{aligned} \tag{2.9}$$

is a positive multiple of the sum

$$\begin{aligned} &K_t(x,y; a,b,c,d; q) \\ &\quad = \sum_{n=0}^{\infty} \frac{(abcdq^{-1})_n(1-abcdq^{2n-1})}{(q)_n(1-abcdq^{-1})} \\ &\quad\quad \cdot \frac{(ab)_n(ac)_n(ad)_n}{(bc)_n(bd)_n(cd)_n} a^{-2n}t^n p_n(x; a,b,c,d; q)p_n(y; a,b,c,d; q), \end{aligned} \tag{2.10}$$

when a, b, c, d are real and their pairwise products have absolute value less than one, which is the case we consider in this paper.

Another notation that we shall use is that of a q-integral defined by

$$\int_a^b f(u)\,d_q u=\int_0^b f(u)\,d_q u-\int_0^a f(u)\,d_q u, \tag{2.11}$$

where

$$\int_0^a f(u)\,d_q u=a(1-q)\sum_{n=0}^{\infty} f(aq^n)q^n$$

whenever the series converges, which will simplify our proofs in §§4 to 6.

3. Quadratic transformations. To derive formula (1.2) we first observe that from a transformation formula due to Bailey [6, (1)] it follows that

(3.1)

$$
{}_5\phi_4\left[\begin{matrix} a, b, c, dq^n, q^{-n} \\ aq/b,\ aq/c,\ aq^{1-n}/d,\ b^2c^2d^2q^{n-2}/a^2 \end{matrix}; q\right]
$$

$$
= \frac{(aq^{2-n}/bcd)_n (a^3q^{3-2n}/b^2c^2d^2)_n}{(a^2q^{2-n}/bcd)_n (a^2q^{3-2n}/b^2c^2d^2)_n}
$$

$$
\cdot {}_{12}\phi_{11}\left[\begin{matrix} a^2q^{1-n}/bcd,\ q\sqrt{\cdot},\ -q\sqrt{\cdot},\ aq^{1-n}/cd, \\ \sqrt{\cdot},\ -\sqrt{\cdot},\ aq/b, \end{matrix}\right.
$$

$$
\begin{matrix} aq^{1-n}/bd,\ aq/bc,\ \sqrt{a},\ -\sqrt{a},\ \sqrt{aq}, \\ aq/c,\ aq^{1-n}/d,\ a^{3/2}q^{2-n}/bcd,\ -a^{3/2}q^{2-n}/bcd,\ a^{3/2}q^{3/2-n}/bcd, \end{matrix}
$$

$$
\left.\begin{matrix} -\sqrt{aq},\ a^3q^{3-n}/b^2c^2d^2,\ q^{-n} \\ -a^{3/2}q^{3/2-n}/bcd,\ bcd/aq,\ a^2q^2/bcd \end{matrix}; q\right].
$$

Letting $n \to \infty$ we get

$$
{}_3\phi_2\left[\begin{matrix} a, b, c \\ aq/b,\ aq/c \end{matrix}; \frac{d}{a}\right] = \frac{(bcd/aq)_\infty}{(bcd/a^2q)_\infty} \lim_{n\to\infty} {}_{12}\phi_{11}[\cdots]. \tag{3.2}
$$

To find the limit on the right side of (3.2) we follow the procedure in Bailey [5, p. 28] of first assuming n to be an odd integer, splitting up the ${}_{12}\phi_{11}$ series into two halves, reversing the second series, and then letting $n \to \infty$. This yields

(3.3)

$$
\lim_{n\to\infty} {}_{12}\phi_{11}[\cdots] = {}_5\phi_4\left[\begin{matrix} \sqrt{a},\ -\sqrt{a},\ \sqrt{aq},\ -\sqrt{aq},\ aq/bc \\ aq/b,\ aq/c,\ bcd/aq,\ a^2q^2/bcd \end{matrix}; q\right]
$$

$$
- \frac{bcd}{a^2q} \frac{(bcd/a^2)_\infty (cd/a)_\infty (bd/a)_\infty (aq/bc)_\infty (a)_\infty}{(d/a)_\infty (aq/b)_\infty (aq/c)_\infty (bcd/aq)_\infty (a^2q^2/bcd)_\infty}
$$

$$
{}_5\phi_4\left[\begin{matrix} d/a,\ bcd/qa^{3/2},\ -bcd/qa^{3/2},\ bcd/q^{1/2}a^{3/2},\ -bcd/q^{1/2}a^{3/2} \\ cd/a,\ bd/a,\ bcd/a^2,\ b^2c^2d^2/a^3q^2 \end{matrix}; q\right].
$$

Replacing d by a^2xq/bc in (3.2) and (3.3) we immediately get (1.2). With Bailey's formula (3.1) known for over 40 years (his paper was accepted for Proceedings of London Mathematical Society in 1943) it is surprising how such a simple derivation of (1.2) could have escaped notice for so long.

Bailey even had a formula [6, (2)] which gives an analogue of (3.1) in the form

(3.4)

$$
{}_7\phi_6\left[\begin{matrix} a,\ q\sqrt{a}\,,\ -q\sqrt{a}\,,\ b,\ c,\ dq^n,\ q^{-n} \\ \sqrt{a}\,,\ -\sqrt{a}\,,\ aq/b,\ aq/c,\ aq^{1-n}/d,\ b^2c^2d^2q^n/a^2 \end{matrix};q\right]
$$

$$
=\frac{(a^3q^{2-2n}/b^2c^2d^2)_{n-1}(aq^{-n}/bcd)_n(1-a^3q/b^2c^2d^2)}{(a^2q^{2-n}/bcd)_n(a^2q^{1-2n}/b^2c^2d^2)_n}
$$

$$
\cdot\,{}_{12}\phi_{11}\left[\begin{matrix} a^2q^{1-n}/bcd,\ q\sqrt{\cdot}\,,\ -q\sqrt{\cdot}\,,\ aq^{1-n}/cd,\ aq^{1-n}/bd,\ aq/bc, \\ \sqrt{\cdot}\,,\ -\sqrt{\cdot}\,,\ aq/b,\ aq/c,\ aq^{1-n}/d, \\ \sqrt{aq}\,,\ -\sqrt{aq}\,,\ q\sqrt{a}\,,\ -q\sqrt{a}\,, \\ a^{3/2}q^{3/2-n}/bcd,\ -a^{3/2}q^{3/2-n}/bcd,\ a^{3/2}q^{1-n}/bcd,\ -a^{3/2}q^{1-n}/bcd, \\ a^3q^{1-n}/b^2c^2d^2,\ q^{-n} \\ bcdq/a,\ a^2q^2/bcd \end{matrix};q\right].
$$

Taking the limit $n\to\infty$ as described above and replacing d by $a^{3/2}q^{1/2}t/bc$ we obtain for the ${}_5\phi_4$ in (1.3) that

(3.5)

$$
{}_5\phi_4\left[\begin{matrix} a,\ q\sqrt{a}\,,\ -q\sqrt{a}\,,\ b,\ c \\ \sqrt{a}\,,\ -\sqrt{a}\,,\ aq/b,aq/c \end{matrix};\frac{t\sqrt{aq}}{bc}\right]
$$

$$
=\frac{(1-t^2)(tq\sqrt{aq})_\infty}{(t/\sqrt{aq})_\infty}\,{}_5\phi_4\left[\begin{matrix} \sqrt{aq}\,,\ -\sqrt{aq}\,,\ q\sqrt{a}\,,\ -q\sqrt{a}\,,\ aq/bc \\ aq/b,\ aq/c,\ t^{-1}q\sqrt{aq}\,,\ tq\sqrt{aq} \end{matrix};q\right]
$$

$$
+\frac{(aq)_\infty(aq/bc)_\infty(\sqrt{aq}\,t/b)_\infty(\sqrt{aq}\,t/c)_\infty}{(aq/b)_\infty(aq/c)_\infty(\sqrt{aq}\,t/bc)_\infty(t^{-1}\sqrt{aq}\,)_\infty}
$$

$$
\cdot\,{}_5\phi_4\left[\begin{matrix} t,\ -t,\ t\sqrt{q}\,,\ -t\sqrt{q}\,,\ \sqrt{aq}\,t/bc \\ \sqrt{aq}\,t/b,\ \sqrt{aq}\,t/c,\ qt^2,\ t\sqrt{q/a} \end{matrix};q\right].
$$

It should be noted that by replacing a, b, c in (3.5) by q^a, q^b, q^c and letting $q\to 1-$ we obtain the quadratic transformation [5, Ex. 6, p. 97]. Similarly, it can be shown that the quadratic transformation [5, Ex. 4(iv) p. 97] is a limit of (1.2).

In terms of the q-integral notation formula (1.2) is equivalent to

(3.6)
$$
{}_3\phi_2\left[\begin{matrix} a,\ b,\ c \\ aq/b,\ aq/c \end{matrix};\frac{aqx}{bc}\right]
$$

$$
=\frac{(a)_\infty(aq/bc)_\infty}{s(1-q)(q)_\infty(aq/b)_\infty(aq/c)_\infty(q/x)_\infty(x)_\infty}
$$

$$
\cdot\int_{xs}^{s}\frac{(qu/xs)_\infty(qu/s)_\infty(aqu/bs)_\infty(aqu/cs)_\infty(axu/s)_\infty}{(\sqrt{a}\,u/s)_\infty(-\sqrt{a}\,u/s)_\infty(\sqrt{aq}\,u/s)_\infty(-\sqrt{aq}\,u/s)_\infty(aqu/bcs)_\infty}\,d_qu
$$

and (3.5) is equivalent to

(3.7)

$$
{}_5\phi_4\left[\begin{matrix} a, q\sqrt{a}, -q\sqrt{a}, b, c \\ \sqrt{a}, -\sqrt{a}, aq/b, aq/c \end{matrix}; \frac{t\sqrt{aq}}{bc}\right]
$$

$$
= \frac{1-t^2}{s(1-q)} \frac{(aq)_\infty (aq/bc)_\infty}{(q)_\infty (aq/b)_\infty (aq/c)_\infty (t/\sqrt{aq})_\infty (t^{-1}q\sqrt{aq})_\infty}
$$

$$
\cdot \int_{st/\sqrt{aq}}^{s} \frac{(q\sqrt{aq}\,u/st)_\infty (qu/s)_\infty (aqu/bs)_\infty (aqu/cs)_\infty (tq\sqrt{aq}\,u/s)_\infty}{(\sqrt{aq}\,u/s)_\infty (-\sqrt{aq}\,u/s)_\infty (q\sqrt{a}\,u/s)_\infty (-q\sqrt{a}\,u/s)_\infty (aqu/bcs)_\infty} d_q u
$$

where s is arbitrary.

It should also be observed that by replacing c by cq^{-n} in (3.1) and (3.4), letting $n \to \infty$, and then setting $d = aq/cx$, we obtain the quadratic transformations

(3.8)

$$
{}_2\phi_1\left[\begin{matrix} a, b \\ aq/b \end{matrix}; \frac{qx}{b^2}\right] = \frac{(qx/b)_\infty (aqx^2/b^2)_\infty}{(aqx/b)_\infty (qx^2/b^2)_\infty}
$$

$$
\cdot {}_8\phi_7\left[\begin{matrix} ax/b, q\sqrt{\cdot}, -q\sqrt{\cdot}, x, \sqrt{a}, -\sqrt{a}, \sqrt{aq}, -\sqrt{aq} \\ \sqrt{\cdot}, -\sqrt{\cdot}, aq/b, \sqrt{a}\,qx/b, -\sqrt{a}\,qx/b, \sqrt{aq}\,x/b, -\sqrt{aq}\,x/b \end{matrix}; \frac{qx}{b^2}\right]
$$

and

(3.9)

$$
{}_4\phi_3\left[\begin{matrix} a, q\sqrt{a}, -q\sqrt{a}, b \\ \sqrt{a}, -\sqrt{a}, aq/b \end{matrix}; \frac{x}{b^2 q}\right]
$$

$$
= \frac{(ax^2/b^2)_\infty (x/bq)_\infty}{(aqx/b)_\infty (x^2/b^2 q)_\infty}
$$

$$
\cdot {}_8\phi_7\left[\begin{matrix} ax/b, q\sqrt{\cdot}, -q\sqrt{\cdot}, x, \sqrt{aq}, -\sqrt{aq}, q\sqrt{a}, -q\sqrt{a} \\ \sqrt{\cdot}, -\sqrt{\cdot}, aq/b, \sqrt{aq}\,x/b, -\sqrt{aq}\,x/b, \sqrt{a}\,x/b, -\sqrt{a}\,x/b \end{matrix}; \frac{x}{b^2 q}\right]
$$

provided that $|qx/b^2| < 1$ in (3.8) and $|x/b^2 q| < 1$ in (3.9) whenever the series do not terminate.

Our derivation of (1.4) and a more general transformation depends on the use of the Sears [14] summation formula

(3.10)

$$
\frac{(e)_\infty (f)_\infty}{(a)_\infty (b)_\infty (c)_\infty} {}_3\phi_2\left[\begin{matrix} a, b, c \\ e, f \end{matrix}; q\right]
$$

$$
- \frac{q}{e} \frac{(q^2/e)_\infty (fq/e)_\infty}{(aq/e)_\infty (bq/e)_\infty (cq/e)_\infty} {}_3\phi_2\left[\begin{matrix} aq/e, bq/e, cq/e \\ q^2/e, fq/e \end{matrix}; q\right]
$$

$$
= \frac{(e)_\infty (q/e)_\infty (f/a)_\infty (f/b)_\infty (f/c)_\infty}{(a)_\infty (b)_\infty (c)_\infty (aq/e)_\infty (bq/e)_\infty (cq/e)_\infty}
$$

where $abcq=ef$. Set $e=-q$, $b=-c$, replace a by aq^r, $r=0,1,2,\cdots$, multiply both sides of (3.10) by

$$\frac{(x^2;\,q^2)_r(y^2;\,q^2)_r}{(-q;\,q)_r(x^2y^2b^2;\,q^2)_r}b^{2r}q^r,$$

use $(a^2;\,q^2)_r=(a;\,q)_r(-a;\,q)_r$, and sum over r from 0 to ∞ to get

(3.11)

$$\frac{(-1)_\infty(-q)_\infty(-ab)_\infty(ab)_\infty(b^2)_\infty}{(a)_\infty(b)_\infty(-b)_\infty(-a)_\infty(b)_\infty(-b)_\infty}\,{}_3\phi_2\left[\begin{matrix}a^2,\ x^2,\ y^2\\ a^2b^2,\ x^2y^2b^2\end{matrix};q^2,\ qb^2\right]$$

$$=\frac{(-q)_\infty(ab^2)_\infty}{(a)_\infty(b)_\infty(-b)_\infty}\sum_{j=0}^{\infty}\frac{(a)_j(b)_j(-b)_j}{(q)_j(-q)_j(ab^2)_j}q^j\,{}_3\phi_2\left[\begin{matrix}q^{-2j},\ x^2,\ y^2\\ x^2y^2b^2,\ b^{-2}q^{2-2j}\end{matrix};q^2,\ q^2\right]$$

$$+\frac{(-q)_\infty(-ab^2)_\infty}{(-a)_\infty(-b)_\infty(b)_\infty}\sum_{j=0}^{\infty}\frac{(-a)_j(-b)_j(b)_j}{(q)_j(-q)_j(-ab^2)_j}q^j\,{}_3\phi_2\left[\begin{matrix}q^{-2j},\ x^2,\ y^2\\ x^2y^2b^2,\ b^{-2}q^{2-2j}\end{matrix};q^2,\ q^2\right].$$

Since the ${}_3\phi_2$ series on the r.h.s. can be summed by using [16; (IV.4)]

$$(3.12)\qquad {}_3\phi_2\left[\begin{matrix}a,\ b,\ q^{-n}\\ c,\ ab/cq^{n-1}\end{matrix};q,\ q\right]=\frac{(c/a;\,q)_n(c/b;\,q)_n}{(c;\,q)_n(c/ab;\,q)_n}$$

with the base q replaced by q^2, it follows from (3.11) that

(3.13)

$${}_3\phi_2\left[\begin{matrix}a^2,\ x^2,\ y^2\\ a^2b^2,\ x^2y^2/b^2\end{matrix};q^2,\ qb^2\right]$$

$$=\frac{(-a;\,q)_\infty(ab^2;\,q)_\infty(b^2;\,q^2)_\infty}{(-1;\,q)_\infty(b^2;\,q)_\infty(a^2b^2;\,q^2)_\infty}\,{}_5\phi_4\left[\begin{matrix}a,\ bx,\ -bx,\ by,\ -by\\ -q,\ ab^2,\ bxy,\ -bxy\end{matrix};q,\ q\right]$$

$$+\frac{(a;\,q)_\infty(-ab^2;\,q)_\infty(b^2;\,q^2)_\infty}{(-1;\,q)_\infty(b^2;\,q)_\infty(a^2b^2;\,q^2)_\infty}\,{}_5\phi_4\left[\begin{matrix}-a,\ -bx,\ bx,\ -by,\ by\\ -q,\ -ab^2,\ -bxy,\ bxy\end{matrix};q,\ q\right].$$

Setting $y=ab$, this reduces to

(3.14)

$${}_2\phi_1\left[\begin{matrix}a^2,\ x^2\\ a^2b^4x^2\end{matrix};q^2,\ qb^2\right]$$

$$=\frac{(-a;\,q)_\infty(ab^2;\,q)_\infty(b^2;\,q^2)_\infty}{(-1;\,q)_\infty(b^2;\,q)_\infty(a^2b^2;\,q^2)_\infty}\,{}_4\phi_3\left[\begin{matrix}a,\ bx,\ -bx,\ -ab^2\\ -q,\ ab^2x,\ -ab^2x\end{matrix};q,\ q\right]$$

$$+\frac{(a;\,q)_\infty(-ab^2;\,q)_\infty(b^2;\,q^2)_\infty}{(-1;\,q)_\infty(b^2;\,q)_\infty(a^2b^2;\,q^2)_\infty}\,{}_4\phi_3\left[\begin{matrix}-a,\ -bx,\ bx,\ ab^2\\ -q,\ -ab^2x,\ ab^2x\end{matrix};q,\ q\right].$$

Using Bailey's [5, p. 69] transformation formula, namely,

(3.15)

$$
{}_8\phi_7\left[\begin{matrix} a, q\sqrt{a}, -q\sqrt{a}, b, c, d, e, f \\ \sqrt{a}, -\sqrt{a}, aq/b, aq/c, aq/d, aq/e, aq/f \end{matrix}; \frac{a^2q^2}{bcdef}\right]
$$

$$
= \frac{(aq)_\infty (aq/de)_\infty (aq/df)_\infty (aq/ef)_\infty}{(aq/d)_\infty (aq/e)_\infty (aq/f)_\infty (aq/def)_\infty}\, {}_4\phi_3\left[\begin{matrix} aq/bc, d, e, f \\ aq/b, aq/c, def/a \end{matrix}; q\right]
$$

$$
+ \frac{(aq)_\infty (aq/bc)_\infty (d)_\infty (e)_\infty (f)_\infty (a^2q^2/cdef)_\infty (a^2q^2/bdef)_\infty}{(aq/b)_\infty (aq/c)_\infty (aq/d)_\infty (aq/e)_\infty (aq/f)_\infty (a^2q^2/bcdef)_\infty (def/aq)_\infty}
$$

$$
\cdot {}_4\phi_3\left[\begin{matrix} aq/de, aq/df, aq/ef, a^2q^2/bcdef \\ aq^2/def, a^2q^2/bdef, a^2q^2/cdef \end{matrix}; q\right],
$$

we obtain

(3.16)
$$
{}_2\phi_1\left[\begin{matrix} a^2, x^2 \\ a^2b^4x^2 \end{matrix}; q^2, qb^2\right]
$$

$$
= \frac{(b^2x^2; q)_\infty (abx; q)_\infty (-abx; q)_\infty (ab^2; q)_\infty (b^2; q^2)_\infty}{(ab^2x^2; q)_\infty (bx; q)_\infty (-bx; q)_\infty (b^2; q)_\infty (a^2b^2; q^2)_\infty}
$$

$$
\cdot {}_8\phi_7\left[\begin{matrix} ab^2x^2q^{-1}, q\sqrt{\cdot}, -q\sqrt{\cdot}, a, x, -x, bx, -bx \\ \sqrt{\cdot}, -\sqrt{\cdot}, b^2x^2, ab^2x, -ab^2x, abx, -abx \end{matrix}; q, ab^2\right].
$$

By Heine's second transformation [2],

(3.17)
$$
{}_2\phi_1\left[\begin{matrix} a^2, x^2 \\ a^2b^4x^2 \end{matrix}; q^2, qb^2\right]
$$

$$
= \frac{(a^2b^2q; q^2)_\infty (b^4x^2; q^2)_\infty}{(b^2q; q^2)_\infty (a^2q^4x^2; q^2)_\infty}\, {}_2\phi_1\left[\begin{matrix} a^2, qb^{-2} \\ a^2b^2q \end{matrix}; q^2, b^4x^2\right]
$$

and hence by using (3.17) and (3.16) and replacing b by $\sqrt{q}/b$ we get formula (1.4). If $a = \pm q^{-n}$, $n = 0, 1, \cdots$, then one of the terms on the r.h.s. of (3.14) drops out and, by (3.17), we get a formula equivalent with Verma's formula (2.5) in [17].

The only tools that we need to derive q-analogues of Watson's (1.5) and Whipple's (1.6) summation formulas are the q-analogue of Dixon's formula [16, (IV.6)]

(3.18)
$$
{}_4\phi_3\left[\begin{matrix} A, -q\sqrt{A}, B, C \\ -\sqrt{A}, Aq/B, Aq/C \end{matrix}; \frac{q\sqrt{A}}{BC}\right]
$$

$$
= \frac{(Aq)_\infty (q\sqrt{A}/C)_\infty (q\sqrt{A}/B)_\infty (Aq/BC)_\infty}{(Aq/B)_\infty (Aq/C)_\infty (q\sqrt{A})_\infty (q\sqrt{A}/BC)_\infty}, \qquad |q\sqrt{A}/BC| < 1,
$$

and the following limit case of Jackson's transformation formula [16, (3.4.2.4)]

(3.19)

$$
{}_8\phi_7\left[\begin{matrix} A, q\sqrt{A}, -q\sqrt{A}, B, C, D, E, F \\ \sqrt{A}, -\sqrt{A}, Aq/B, Aq/C, Aq/D, Aq/E, Aq/F \end{matrix}; \frac{A^2q^2}{BCDEF}\right]
$$

$$
= \frac{(Aq)_\infty (Aq/EF)_\infty (A^2q^2/BCDE)_\infty (A^2q^2/BCDF)_\infty}{(Aq/E)_\infty (Aq/F)_\infty (A^2q^2/BCD)_\infty (A^2q^2/BCDEF)_\infty}
$$

$$
\cdot {}_8\phi_7\left[\begin{matrix} A^2q/BCD, q\sqrt{\cdot}, -q\sqrt{\cdot}, Aq/CD, Aq/BD, Aq/BC, E, F \\ \sqrt{\cdot}, -\sqrt{\cdot}, Aq/B, Aq/C, Aq/D, A^2q^2/BCDE, A^2q^2/BCDF \end{matrix}; \frac{Aq}{EF}\right],
$$

provided that $|A^2q^2/BCDEF|<1$ and $|Aq/EF|<1$ whenever the series do not terminate.

If we now set $A=-F=-a$, $B=-C=b$, and $D=-E=c$, then the l.h.s. of (3.19) reduces to the ${}_4\phi_3$ series

$$
{}_4\phi_3\left[\begin{matrix} a^2, -aq^2, b^2, c^2 \\ -a, a^2q^2/b, a^2q^2/c^2 \end{matrix}; q^2, \frac{aq^2}{b^2c^2}\right]
$$

which, by (3.18), sums to

$$
\frac{(a^2q^2;\, q^2)_\infty (aq^2/c^2;\, q^2)_\infty (aq^2/b^2;\, q^2)_\infty (a^2q^2/b^2c^2;\, q^2)_\infty}{(a^2q^2/b^2;\, q^2)_\infty (a^2q^2/c^2;\, q^2)_\infty (aq^2;\, q^2)_\infty (aq^2/b^2c^2;\, q^2)_\infty}.
$$

Hence it follows from (3.19) that

(3.20)
$$
{}_8\phi_7\left[\begin{matrix} -a^2q/b^2c, q\sqrt{\cdot}, -q\sqrt{\cdot}, aq/bc, -aq/bc, aq/b^2, -c, a \\ \sqrt{\cdot}, -\sqrt{\cdot}, -aq/b, aq/b, -aq/c, a^2q^2/b^2c^2, -aq^2/b^2c \end{matrix}; q, \frac{q}{c}\right]
$$

$$
= \frac{(aq/c;\, q)_\infty (-q;\, q)_\infty (-a^2q^2/b^2c;\, q)_\infty (aq^2/b^2c^2;\, q)_\infty}{(-aq;\, q)_\infty (q/c;\, q)_\infty (a^2q^2/b^2c^2;\, q)_\infty (-aq^2/b^2c;\, q)_\infty}
$$

$$
\cdot \frac{(a^2q^2;\, q^2)_\infty (aq^2/b^2;\, q^2)_\infty (aq^2/c^2;\, q^2)_\infty (a^2q^2/b^2c^2;\, q^2)_\infty}{(a^2q^2/b^2;\, q^2)_\infty (a^2q^2/c^2;\, q^2)_\infty (aq^2;\, q^2)_\infty (aq^2/b^2c^2;\, q^2)_\infty}.
$$

Replacing aq/b^2, aq/bc, and c by b, c, and $\sqrt{abq}/c$, respectively, we obtain the following q-analogue of (1.5):

(3.21)

$$
{}_8\phi_7\left[\begin{matrix} -c\sqrt{ab/q}, q\sqrt{\cdot}, -q\sqrt{\cdot}, a, b, c, -c, -c^{-1}\sqrt{abq} \\ \sqrt{\cdot}, -\sqrt{\cdot}, -c\sqrt{bq/a}, -c\sqrt{aq/b}, -\sqrt{abq}, \sqrt{abq}, c^2 \end{matrix}; q, \frac{c\sqrt{q}}{\sqrt{ab}}\right]
$$

$$
= \frac{(c\sqrt{aq/b}\,;\, q)_\infty (-q;\, q)_\infty (-c\sqrt{abq}\,;\, q)_\infty (c^2/a;\, q)_\infty}{(-aq;\, q)_\infty (c\sqrt{q/ab}\,;\, q)_\infty (c^2;\, q)_\infty (-c\sqrt{bq/a}\,;\, q)_\infty}
$$

$$
\cdot \frac{(a^2q^2;\, q^2)_\infty (bq;\, q^2)_\infty (qc^2/b^2;\, q^2)_\infty (c^2;\, q^2)_\infty}{(abq;\, q^2)_\infty (ac^2q/b^2;\, q^2)_\infty (aq^2;\, q^2)_\infty (c^2/a;\, q^2)_\infty}, \qquad \left|\frac{c^2q}{ab}\right|<1.
$$

To obtain a q-analogue of Whipple's summation formula (1.6) we apply (3.19) to the ${}_8\phi_7$ in (3.20) by identifying A, B, C, D, E, and F with $-a^2q/b^2c$, $-aq/bc$, aq/b^2, a, aq/bc and $-c$, respectively. This yields a summation formula for an ${}_8\phi_7$ which on replacing a, b, c by bc/q, a, c/b, respectively, gives the desired formula

$$(3.22)\quad {}_8\phi_7\left[\begin{matrix} -b, q\sqrt{\cdot}, -q\sqrt{\cdot}, a, q/a, b, -bq/c, -c/b \\ \sqrt{\cdot}, -\sqrt{\cdot}, -bq/a, -ab, -q, c, b^2q/c \end{matrix}; q, b\right]$$

$$= \frac{(-c;q)_\infty(-bq;q)_\infty(ab;q)_\infty(b^2q/ac;q)_\infty}{(b;q)_\infty(b^2q/c;q)_\infty(-ac;q)_\infty(-bq/a;q)_\infty}$$

$$\cdot \frac{(a^2c^2;q^2)_\infty(cq/a;q^2)_\infty(ab^2q/c;q^2)_\infty(b^2;q^2)_\infty}{(c^2;q^2)_\infty(a^2b^2;q^2)_\infty(acq;q^2)_\infty(b^2q/ac;q^2)_\infty}, \qquad |b|<1.$$

Additional summation formulas for ${}_8\phi_7$ series follow by applying the transformation formula (3.19) to formulas (3.21) and (3.22).

It should be noted that if we set a or b in (3.21) equal to q^{-n}, $n=0,1,\cdots$, then we obtain Theorem 1 in Andrews [3] by means of Watson's [16, (3.4.1.5)] transformation formula

$$(3.23)\quad {}_8\phi_7\left[\begin{matrix} a, q\sqrt{a}, -q\sqrt{a}, b, c, d, e, q^{-n} \\ \sqrt{a}, -\sqrt{a}, aq/b, aq/c, aq/d, aq/e, aq^{n+1} \end{matrix}; \frac{aq^{n+2}}{bcde}\right]$$

$$= \frac{(aq)_n(aq/de)_n}{(aq/d)_n(aq/e)_n}\,{}_4\phi_3\left[\begin{matrix} aq/bc, d, e, q^{-n} \\ de/aq^n, aq/b, aq/c \end{matrix}; q\right].$$

Similarly, Theorem 2 in Andrews [3] follows by applying (3.23) to the case $a=q^{-n}$ of (3.22).

In the next section we will need to use the fact that Sears' summation formula (3.10) is equivalent to the formula

$$(3.24)\quad \int_a^b \frac{(qu/a)_\infty(qu/b)_\infty(cu)_\infty}{(du)_\infty(eu)_\infty(fu)_\infty}\,d_qu$$

$$= \frac{b(1-q)(q)_\infty(bq/a)_\infty(a/b)_\infty(c/d)_\infty(c/e)_\infty(c/f)_\infty}{(ad)_\infty(ae)_\infty(af)_\infty(bd)_\infty(be)_\infty(bf)_\infty},$$

where $c=abdef$, which is also equivalent to the q-beta formula (1.3) in Al-Salam and Verma [1]. Likewise, Bailey's transformation formula (3.15) is equivalent to

$$(3.25)\quad \int_a^b d_qu \frac{(qu/a)_\infty(qu/b)_\infty(cu)_\infty(du)_\infty}{(eu)_\infty(fu)_\infty(gu)_\infty(hu)_\infty}$$

$$= \frac{b(1-q)(q)_\infty(bq/a)_\infty(a/b)_\infty(cd/eh)_\infty(cd/fh)_\infty(cd/gh)_\infty(bc)_\infty(bd)_\infty}{(ae)_\infty(af)_\infty(ag)_\infty(be)_\infty(bf)_\infty(bg)_\infty(bh)_\infty(bcd/h)_\infty}$$

$$\cdot {}_8\phi_7\left[\begin{matrix} bcd/hq, q\sqrt{\cdot}, -q\sqrt{\cdot}, be, bf, bg, c/h, d/h \\ \sqrt{\cdot}, -\sqrt{\cdot}, cd/eh, cd/fh, cd/gh, bd, bc \end{matrix}; cd/befg\right],$$

where $cd=abefgh$. It should also be observed that formula (IV.15) in [16] for the sum of two ${}_8\phi_7$ series is equivalent to

(3.26)

$$\int_a^b \frac{(qu/a)_\infty (qu/b)_\infty (u/\sqrt{a})_\infty (-u/\sqrt{a})_\infty (qu/c)_\infty (qu/d)_\infty (qu/e)_\infty (qu/f)_\infty}{(u)_\infty (qu/\sqrt{a})_\infty (-qu/\sqrt{a})_\infty (bu/a)_\infty (cu/a)_\infty (du/a)_\infty (eu/a)_\infty (fu/a)_\infty} d_q u$$

$$= \frac{a(q-1)(q)_\infty (aq/b)_\infty (b/a)_\infty (aq/cd)_\infty}{(b)_\infty (c)_\infty (d)_\infty (e)_\infty (f)_\infty}$$

$$\cdot \frac{(aq/ce)_\infty (aq/de)_\infty (ae/cf)_\infty (aq/df)_\infty (aq/ef)_\infty}{(bc/a)_\infty (bd/a)_\infty (be/a)_\infty (bf/a)_\infty},$$

where $a^2q=bcdef$, which has (3.24) as a limit case.

As usual, in (3.26) and other formulas it is assumed that no zero factors appear in the denominator.

4. Some generating functions for q-Wilson polynomials. Here, as an introduction to the method that we will apply to the Poisson kernel, we first consider the simpler sum

$$G_t(x;\, a,b,c,d;\, q) = \sum_{n=0}^{\infty} \frac{(abcd/q)_n (ad)_n}{(q)_n (bc)_n} \left(\frac{t}{a}\right)^n p_n(x;\, a,b,c,d;\, q), \tag{4.1}$$

which, by the asymptotic formula for the q-Wilson polynomials in Ismail and Wilson [12, (1.13)], converges for $|t|<1$ when $-1<x<1$. Unfortunately, if the ${}_4\phi_3$ series representation for $p_n(x;\, a,b,c,d;\, q)$ in (2.5) is used in (4.1), then we cannot change the order of summation and sum over n since, due to the term q^{-n} in the numerator of the ${}_4\phi_3$ series, this leads to a divergent series when $t\neq 0$. This divergence problem also keeps us from applying the product formulas in Gasper and Rahman [10] to the Poisson kernel (2.9). To overcome this difficulty in (4.1) we can either use (3.23) to write the q-Wilson polynomial in terms of a ${}_8\phi_7$ in which q^{-n} also appears in the denominator or write it as an appropriate sum of ${}_4\phi_3$ or ${}_8\phi_7$ series. Here we shall employ the q-integral representation

(4.2)

$$p_n(x;\, a,b,c,d;\, q)$$

$$= A(\theta) \frac{(bc)_n}{(ad)_n} \int_{qe^{i\theta}/d}^{qe^{-i\theta}/d} \frac{(abcdu/q)_\infty (due^{i\theta})_\infty (due^{-i\theta})_\infty (q/u)_n}{(cdu/q)_\infty (bdu/q)_\infty (adu/q)_\infty (abcdu/q)_n} \left(\frac{adu}{q}\right)^n d_q u,$$

where

$$A(\theta) = \frac{2id}{q(1-q)(q)_\infty (ab)_\infty (ac)_\infty (bc)_\infty (de^{i\theta})_\infty (de^{-i\theta})_\infty w(x;\, a,b,c,d;\, q)}$$

with $x=\cos\theta$, which follows from formula (3.3) in Al-Salam and Verma [1] and is equivalent to a representation as a sum of two ${}_4\phi_3$ series.

After substituting (4.2) into (4.1) we can now sum over n to obtain

$$(4.3)\qquad G_t(x;\,a,b,c,d;\,q)=A(\theta)\int_{qe^{i\theta}/d}^{qe^{-i\theta}/d}\frac{(abcdu/q)_\infty(due^{i\theta})_\infty(due^{-i\theta})_\infty}{(adu/q)_\infty(bdu/q)_\infty(cdu/q)_\infty}\cdot{}_2\phi_1\left[\begin{matrix}abcd/q,\ q/u\\ abcdu/q\end{matrix};\frac{dtu}{q}\right]d_qu.$$

From (3.6),

$$(4.4)\qquad {}_2\phi_1\left[\begin{matrix}abcd/q,\ q/u\\ abcdu/q\end{matrix};\frac{dtu}{q}\right]=\frac{(abcd/q)_\infty(\alpha u/q)_\infty}{(1-q)(q)_\infty(abcdu/q)_\infty(\alpha)_\infty(\alpha q/dt)_\infty(dt/\alpha)_\infty}\cdot\int_{dt/\alpha}^{1}\frac{(\alpha qv/dt)_\infty(qv)_\infty(abcduv/q)_\infty(\alpha dtv/q)_\infty}{(\alpha v/\sqrt{q})_\infty(-\alpha v/\sqrt{q})_\infty(-\alpha v)_\infty(\alpha uv/q)_\infty}d_qv,$$

with $\alpha=(abcd)^{1/2}$. Using (4.4) in (4.3) and changing the order of integration gives

$$(4.5)\qquad G_t(x;\,a,b,c,d;\,q)=\frac{A(\theta)(abcd/q)_\infty}{(1-q)(q)_\infty(\alpha)_\infty(\alpha q/dt)_\infty(dt/\alpha)_\infty}\cdot\int_{dt/\alpha}^{1}d_qv\frac{(\alpha qv/dt)_\infty(qv)_\infty(\alpha dtv/q)_\infty}{(\alpha v/\sqrt{q})_\infty(-\alpha v/\sqrt{q})_\infty(-\alpha v)_\infty}\cdot\int_{qe^{i\theta}/d}^{qe^{-i\theta}/d}d_qu\frac{(\alpha u/q)_\infty(abcduv/q)_\infty(due^{i\theta})_\infty(due^{-i\theta})_\infty}{(adu/q)_\infty(bdu/q)_\infty(cdu/q)_\infty(\alpha uv/q)_\infty}.$$

At this point if in addition to the assumption that $-1<a,\ b,\ c,\ d<1$, we assume $ad=bc$ then the last integral in (4.5) can be evaluated via (3.24) to give

$$(4.6)\qquad G_t(x;\,a,b,c,bc/a;\,q)=\frac{(b^2c^2/q)_\infty(ae^{i\theta})_\infty(ae^{-i\theta})_\infty}{(1-q)(q)_\infty(bc)_\infty(aq/t)_\infty(t/a)_\infty(ab)_\infty(ac)_\infty}\cdot\int_{t/a}^{1}d_qv\frac{(aqv/t)_\infty(qv)_\infty(b^2c^2tv/aq)_\infty(abv)_\infty(acv)_\infty}{(bcv/\sqrt{q})_\infty(-bcv/\sqrt{q})_\infty(-bcv)_\infty(ave^{i\theta})_\infty(ave^{-i\theta})_\infty}$$

$$=\frac{(b^2c^2t/aq)_\infty}{(t/a)_\infty}\,{}_5\phi_4\left[\begin{matrix}bc/\sqrt{q},\ -bc/\sqrt{q},\ -bc,\ ae^{i\theta},\ ae^{-i\theta}\\ ab,\ ac,\ aq/t,\ b^2c^2t/aq\end{matrix};q\right]+\frac{(bct/a)_\infty(b^2c^2/q)_\infty(bt)_\infty(ct)_\infty(ae^{i\theta})_\infty(ae^{-i\theta})_\infty}{(a/t)_\infty(bc)_\infty(ab)_\infty(ac)_\infty(te^{i\theta})_\infty(te^{-i\theta})_\infty}\cdot{}_5\phi_4\left[\begin{matrix}bct/a\sqrt{q},\ -bct/a\sqrt{q},\ -bct/a,\ te^{i\theta},\ te^{-i\theta}\\ bt,\ ct,\ tq/a,\ b^2c^2t^2/a^2q\end{matrix};q\right].$$

Since

$$(4.7)\quad P_n^{(\alpha,\beta)}(x\,|\,q)=\frac{(q^{\alpha+1})_n}{(q)_n}p_n(x;\ q^{\alpha/2+1/4},\ q^{\alpha/2+3/4},\ -q^{\beta/2+1/4},\ -q^{\beta/2+3/4};\ q)$$

by (2.3), for the continuous q-Jacobi polynomials the sum of two ${}_5\phi_4$ series in (4.6) reduces to a sum of two ${}_4\phi_3$ series and it then follows from (3.15) that

(4.8)

$$G_t(x;\ a,b,c,d;\ q)=\frac{(bt/\sqrt{q})_\infty(bc^2t/\sqrt{q})_\infty(-bcte^{i\theta}/\sqrt{q})_\infty(-bcte^{-i\theta}/\sqrt{q})_\infty}{(-ct)_\infty(-b^2ct/q)_\infty(te^{i\theta})_\infty(te^{-i\theta})_\infty}$$

$$\cdot\,{}_8\phi_7\left[\begin{matrix}-b^2ct/q^2,\ q\sqrt{\cdot},\ -q\sqrt{\cdot},\ -ct/\sqrt{q},\\ \sqrt{\cdot},\ -\sqrt{\cdot},\ b^2/\sqrt{q},\ bc^2t/\sqrt{q},\end{matrix}\right.$$

$$\left.\cdot\,\begin{matrix}-b/c\sqrt{q},\ -bc/\sqrt{q},\ be^{i\theta}/\sqrt{q},\ be^{-i\theta}/\sqrt{q}\\ bt/\sqrt{q},\ -bcte^{-i\theta}/\sqrt{q},\ -bcte^{i\theta}/\sqrt{q}\end{matrix}\ ;\ -ct\sqrt{q}\right]$$

with

$$(4.9)\qquad a=q^{\alpha/2+1/4},\quad b=q^{\alpha/2+3/4},\quad c=-q^{\beta/2+1/4},\quad d=-q^{\beta/2+3/4}.$$

If $\alpha=\beta$, then the r.h.s. of (4.8) degenerates to

$$\frac{(te^{i\theta}q^{\beta+1/2})_\infty(te^{-i\theta}q^{\beta+1/2})_\infty}{(te^{i\theta})_\infty(te^{-i\theta})_\infty},$$

giving a well-known generating function for the continuous q-ultraspherical polynomials.

Now consider

$$(4.10)\quad K_t(x;\ a,b,c,d;\ q)=\sum_{n=0}^{\infty}\frac{(abcd/q)_n(1-abcdq^{2n-1})}{(q)_n(1-abcdq^{-1})}$$

$$\cdot\,\frac{(ab)_n(ac)_n(ad)_n}{(bc)_n(bd)_n(cd)_n}a^{-2n}t^n p_n(x;\ a,b,c,d;\ q),$$

which is the kernel (2.10) with $p_n(y;\ a,b,c,d;\ q)$ replaced by 1, and converges for $|t|<|a|$ when $-1<x<1$. By employing the above method with (3.7) replaced by (3.6) and using the fact that (via [15, (8.3)])

$$p_n(x;\ a,b,c,d;\ q)=\frac{(bd)_n(cd)_n}{(ab)_n(ac)_n}\left(\frac{a}{d}\right)^n p_n(x;\ d,c,b,a;\ q)$$

we find that

(4.11) $K_t(x;\, a,b,ad/b,d;\, q)$

$$=\frac{(1-t^2)(adqt)_\infty}{(t/ad)_\infty}\,{}_5\phi_4\left[\begin{matrix} ad\sqrt{q}\,,\ -ad\sqrt{q}\,,\ -ad,\ de^{i\theta},\ de^{-i\theta} \\ bd,\ ad^2/b,\ adqt,\ adq/t \end{matrix};q\right]$$

$$+\frac{(t)_\infty(bt/a)_\infty(dt/b)_\infty(a^2d^2)_\infty(de^{i\theta})_\infty(de^{-i\theta})_\infty}{(ad/t)_\infty(ad)_\infty(bd)_\infty(ad^2/b)_\infty(te^{i\theta}/a)_\infty(te^{-i\theta}/a)_\infty}$$

$$\cdot\,{}_5\phi_4\left[\begin{matrix} t\sqrt{q}\,,\ -t\sqrt{q}\,,\ -t,\ te^{i\theta}/a,\ te^{-i\theta}/a \\ bt/a,\ dt/b,\ qt/ad,\ qt^2 \end{matrix};q\right].$$

As above, for the continuous q-Jacobi polynomials it follows from (4.11) and (3.15) that

(4.12)

$$K_t(x;\, a,b,c,d;\, q)=\frac{(1-t^2)(adqt)_\infty(dt/a)_\infty(-dte^{i\theta})_\infty(-dte^{-i\theta})_\infty}{(-t)_\infty(-d^2t)_\infty(te^{i\theta}/a)_\infty(te^{-i\theta}/a)_\infty}$$

$$\cdot\,{}_8\phi_7\left[\begin{matrix} -d^2t/q,\ q\sqrt{\cdot}\,,\ -q\sqrt{\cdot}\,,\ -t\sqrt{q}\,,\ -d/aq,\ -ad,\ de^{i\theta},\ de^{-i\theta} \\ \sqrt{\cdot}\,,\ -\sqrt{\cdot}\,,\ d^2/\sqrt{q}\,,\ adqt,\ dt/a,\ -dte^{-i\theta},\ -dte^{i\theta} \end{matrix};-tq^{1/2}\right]$$

when (4.9) holds. The transformation formula (3.19) can be applied to the ${}_8\phi_7$ series in (4.8) and (4.12) to give cases in which these generating functions are positive, but we shall not do so here, since we are primarily concerned with the positivity of the Poisson kernel.

5. The kernel $K_t(x, y; a, b, c, d; q)$. Let $x=\cos\theta$, $y=\cos\phi$, $0\leqq t<1$, $0\leqq q<1$, and let a, b, c, d be real. Since, by (4.2),

(5.1)

$$p_n(x;\, a,b,c,d;\, q)\,p_n(y;\, a,b,c,d;\, q)$$

$$=B(\theta,\phi)\frac{(cd)_n(bd)_n}{(ab)_n(ac)_n}\int_{qe^{i\theta}/b}^{qe^{-i\theta}/b}d_qu\,\frac{(abcdu/q)_\infty(bue^{i\theta})_\infty(bue^{-i\theta})_\infty}{(abu/q)_\infty(bcu/q)_\infty(bdu/q)_\infty}$$

$$\cdot\int_{qe^{i\phi}/c}^{qe^{-i\phi}/c}d_qv\,\frac{(abcdv/q)_\infty(cve^{i\phi})_\infty(cve^{-i\phi})_\infty}{(acv/q)_\infty(bcv/q)_\infty(cdv/q)_\infty}\,\frac{(q/u)_n(q/v)_n}{(abcdu/q)_n(abcdv/q)_n}\left(\frac{a^2bcuv}{q^2}\right)^n,$$

with

(5.2)

$$B(\theta,\phi)=B(\theta,\phi;\, a,b,c,d;\, q)$$

$$=\frac{-4bc[w(x;\, a,b,c,d;\, q)w(y;\, a,b,c,d;\, q)]^{-1}}{[q(1-q)(q)_\infty(ad)_\infty]^2(ab)_\infty(ac)_\infty(bd)_\infty(cd)_\infty|(be^{i\theta})_\infty(ce^{i\phi})_\infty|^2},$$

it follows from (2.10) that

$$K_t(x,y;\,a,b,c,d;\,q)$$

$$= B(\theta,\phi)\int_{qe^{i\theta}/b}^{qe^{-i\theta}/b} d_q u\,\frac{(abcdu/q)_\infty (bue^{i\theta})_\infty (bue^{-i\theta})_\infty}{(abu/q)_\infty (bcu/q)_\infty (bdu/q)_\infty}$$

$$\cdot \int_{qe^{i\phi}/c}^{qe^{-i\phi}/c} d_q v\,\frac{(abcdv/q)_\infty (cve^{i\phi})_\infty (cve^{-i\phi})_\infty}{(acv/q)_\infty (bcv/q)_\infty (cdv/q)_\infty}$$

$$\cdot {}_6\phi_5\left[\begin{matrix} abcd/q,\ q\sqrt{\cdot},\ -q\sqrt{\cdot},\ q/u,\ q/v,\ ad \\ \sqrt{\cdot},\ -\sqrt{\cdot},\ abcdu/q,\ abcdv/q,\ bc \end{matrix};\ \frac{bcuvt}{q^2}\right].$$

At this point we have to assume that

$$ad = bc, \tag{5.4}$$

so that the above ${}_6\phi_5$ reduces to a ${}_5\phi_4$ to which we can apply formula (3.5) or (3.7). Here we shall apply formula (3.5) since, as we shall see, the first ${}_5\phi_4$ on the r.h.s. of (3.5) leads to a single sum of a terminating ${}_{10}\phi_9$ series. This yields

(5.5)

$$K_t(x,y;\,a,b,c,bc/a;\,q)$$

$$= B(\theta,\phi;\,a,b,c,bc/a;\,q)(1-t^2)\frac{(bcqt)_\infty}{(t/bc)_\infty}$$

$$\cdot \sum_{k=0}^{\infty} \frac{(b^2c^2)_{2k}q^k}{(q)_k (bcqt)_k (bcq/t)_k} \int_{qe^{i\theta}/b}^{qe^{-i\theta}/b} d_q u\,\frac{(b^2c^2uq^{k-1})_\infty (bue^{i\theta})_\infty (bue^{-i\theta})_\infty}{(abu/q)_\infty (bcu/q)_\infty (b^2cu/aq)_\infty}$$

$$\cdot \int_{qe^{i\phi}/c}^{qe^{-i\phi}/c} d_q v\,\frac{(b^2c^2vq^{k-1})_\infty (cve^{i\phi})_\infty (cve^{-i\phi})_\infty (b^2c^2uv/q^2)_\infty}{(acv/q)_\infty (bcv/q)_\infty (bc^2v/aq)_\infty (b^2c^2uvq^{k-2})_\infty}$$

$$+ B(\theta,\phi;\,a,b,c,bc/a;\,q)\frac{(b^2c^2)_\infty}{(bc/t)_\infty} \sum_{k=0}^{\infty} \frac{(t^2)_{2k}q^k}{(q)_k (qt/bc)_k (qt^2)_k}$$

$$\cdot \int_{qe^{i\theta}/b}^{qe^{-i\theta}/b} d_q u\,\frac{(bcutq^{k-1})_\infty (bue^{i\theta})_\infty (bue^{-i\theta})_\infty}{(abu/q)_\infty (bcu/q)_\infty (b^2cu/aq)_\infty}$$

$$\cdot \int_{qe^{i\phi}/c}^{qe^{-i\phi}/c} d_q v\,\frac{(bcvtq^{k-1})_\infty (cve^{i\phi})_\infty (cve^{-i\phi})_\infty (b^2c^2uv/q^2)_\infty}{(acv/q)_\infty (bcv/q)_\infty (bc^2v/aq)_\infty (bcuvtq^{k-2})_\infty}.$$

From (3.15),

(5.6)

$$\int_{qe^{i\phi}/c}^{qe^{-i\phi}/c} d_q v \frac{(b^2c^2vq^{k-1})_\infty (cve^{i\phi})_\infty (cve^{-i\phi})_\infty (b^2c^2uv/q^2)_\infty}{(acv/q)_\infty (bcv/q)_\infty (bc^2v/aq)_\infty (b^2c^2uvq^{k-2})_\infty}$$

$$= \frac{q(1-q)(q)_\infty (b^2c/a)_\infty (ab)_\infty (bc)_\infty (e^{2i\phi})_\infty (e^{-2i\phi})_\infty}{2ci\sin\phi (b^2ce^{-i\phi})_\infty (ae^{i\phi})_\infty (be^{i\phi})_\infty (bce^{i\phi}/a)_\infty (bce^{-i\phi}/a)_\infty}$$

$$\cdot \frac{(b^2ce^{-i\phi}q^k)_\infty (b^2cue^{-i\phi}/q)_\infty}{(ae^{-i\phi})_\infty (be^{-i\phi})_\infty (b^2cue^{-i\phi}q^{k-1})_\infty}$$

$$\cdot {}_8\phi_7\left[\begin{matrix} b^2ce^{-i\phi}/q, q\sqrt{\cdot}, -q\sqrt{\cdot}, bce^{-i\phi}/a, be^{-i\phi}, q/u, ae^{-i\phi}, q^{-k} \\ \sqrt{\cdot}, -\sqrt{\cdot}, ab, bc, b^2cue^{-i\phi}/q, b^2c/a, b^2ce^{-i\phi}q^k \end{matrix}; b^2cue^{i\phi}q^{k-1}\right]$$

$$= \frac{q(1-q)}{2ci}(q)_\infty (ab)_\infty (bc)_\infty (b^2c/a)_\infty (ce^{i\phi})_\infty (ce^{-i\phi})_\infty w(y; a,b,c,bc/a; q)$$

$$\cdot \frac{(b^2cu/aq)_k}{(b^2c/a)_k} {}_4\phi_3\left[\begin{matrix} q^{-k}, q/u, ae^{i\phi}, ae^{-i\phi} \\ ab, bc, a/b^2cuq^{k-2} \end{matrix}; q\right]$$

by (3.23). Hence

(5.7)

$$\int_{qe^{i\theta}/b}^{qe^{-i\theta}/b} d_q u \frac{(b^2c^2uq^{k-1})_\infty (bue^{i\theta})_\infty (bue^{-i\theta})_\infty}{(abu/q)_\infty (bcu/q)_\infty (b^2cu/aq)_\infty} \int_{qe^{i\phi}/c}^{qe^{-i\phi}/c} d_q v \cdots$$

$$= \frac{q(1-q)}{2ci}(q)_\infty (ab)_\infty (bc)_\infty (b^2c/a)_\infty (ce^{i\phi})_\infty (ce^{-i\phi})_\infty w(y; a,b,c,bc/a; q)$$

$$\cdot \sum_{j=0}^{k} \frac{(q)_k (ae^{i\phi})_j (ae^{-i\phi})_j}{(q)_j (q)_{k-j} (ab)_j (bc)_j (b^2c/a)_k} \left(\frac{b^2c}{aq}\right)^j$$

$$\cdot \int_{qe^{i\theta}/b}^{qe^{-i\theta}/b} d_q u \frac{(b^2c^2uq^{k-1})_\infty (bue^{i\theta})_\infty (bue^{-i\theta})_\infty (q/u)_j u^j}{(abu/q)_\infty (bcu/q)_\infty (a^{-1}b^2cuq^{k-j-1})_\infty}$$

$$= -\frac{[q(1-q)(q)_\infty (bc)_\infty]^2 (ab)_\infty (ac)_\infty (b^2c/a)_\infty (bc^2/a)_\infty}{4bc(bc)_k (b^2c/a)_k (bc^2/a)_k}$$

$$\cdot \left|(be^{i\theta})_\infty (ce^{i\phi})_\infty (bca^{-1}e^{i\theta})_k\right|^2 w(x; a,b,c,bc/a; q) w(y; a,b,c,bc/a; q)$$

$$\cdot \sum_{j=0}^{k} \frac{(q^{-k})_j (a/bc^2q^{k-1})_j \left|(ae^{i\phi})_j\right|^2}{(q)_j (ab)_j \left|(ae^{i\theta}/bcq^{k-1})_j\right|^2} q^j$$

$$\cdot {}_4\phi_3\left[\begin{matrix} q^{-j}, b^2c^2q^{k-1}, ce^{i\theta}, ce^{-i\theta} \\ ac, bc, bc^2a^{-1}q^{k-j} \end{matrix}; q\right]$$

by using (3.15) and (3.23) as in (5.6). Consequently, the first term on the r.h.s. of (5.5) gives

$$(5.8)\qquad (1-t^2)\frac{(bcqt)_\infty}{(t/bc)_\infty}\sum_{k=0}^{\infty}\frac{(-bc)_k\left(bc\sqrt{q}\right)_k\left(-bc\sqrt{q}\right)_k\left|\left(bca^{-1}e^{i\theta}\right)_k\right|^2}{(q)_k\left(bc^2/a\right)_k\left(b^2c/a\right)_k(bcqt)_k(bcq/t)_k}q^k$$

$$\cdot\sum_{j=0}^{k}\frac{\left(q^{-k}\right)_j\left(a/bc^2q^{k-1}\right)_j\left|\left(ae^{i\theta}\right)_j\right|^2}{(q)_j(ab)_j\left|\left(ae^{i\theta}/bcq^{k-1}\right)_j\right|^2}q^j$$

$$\cdot\,{}_4\phi_3\left[\begin{matrix}q^{-j},\ b^2c^2q^{k-1},\ ce^{i\theta},\ ce^{-i\theta}\\ ac,\ bc,\ bc^2a^{-1}q^{k-j}\end{matrix};q\right].$$

Transforming the above ${}_4\phi_3$ via [15, (8.3)], the sum over j in (5.8) becomes

(5.9)

$$\sum_{j=0}^{k}\frac{\left(q^{-k}\right)_j\left(b^{-1}c^{-1}q^{1-k}\right)_j\left|\left(ae^{i\phi}\right)_j\right|^2q^j}{(q)_j(bc)_j\left|\left(ae^{i\theta}/bcq^{k-1}\right)_j\right|^2}\,{}_4\phi_3\left[\begin{matrix}q^{-j},\ b^2c^2q^{k-1},\ ae^{i\theta},\ ae^{-i\theta}\\ ab,\ ac,\ bcq^{k-j}\end{matrix};q\right]$$

$$=\sum_{j=0}^{k}\frac{\left(q^{-k}\right)_j\left|\left(ae^{i\phi}\right)_j\right|^2q^j}{(q)_j(bc)_j\left(a^2/bcq^{k-1}\right)_j}\,{}_8\phi_7\left[\begin{matrix}a^2/bcq^k,\ q\sqrt{\cdot},\ -q\sqrt{\cdot},\ a/b^2cq^{k-1},\\ \sqrt{\cdot},\ -\sqrt{\cdot},\ ab,\end{matrix}\right.$$

$$\left.\begin{matrix}a/bc^2q^{k-1},\ ae^{i\theta},\ ae^{-i\theta},\ q^{-j}\\ ac,\ ae^{-i\theta}/bcq^{k-1},\ ae^{i\theta}/bcq^{k-1},\ a^2/bcq^{k-j-1}\end{matrix};bcq^j\right]$$

$$=\sum_{m=0}^{k}\frac{\left(a^2/bcq^k\right)_m\left(q\sqrt{\cdot}\right)_m\left(-q\sqrt{\cdot}\right)_m\left(a/b^2cq^{k-1}\right)_m\left(a/bc^2q^{k-1}\right)_m\left|\left(ae^{i\theta}\right)_m\right|^2}{(q)_m\left(\sqrt{\cdot}\right)_m\left(-\sqrt{\cdot}\right)_m(ab)_m(ac)_m(bc)_m\left(a^2/bcq^{k-1}\right)_{2m}\left|\left(ae^{i\theta}/bcq^{k-1}\right)_m\right|^2}$$

$$\cdot(-1)^mq^{m(m+1)/2}(bc)^m\left(q^{-k}\right)_m\left|\left(ae^{i\phi}\right)_m\right|^2$$

$$\cdot\,{}_3\phi_2\left[\begin{matrix}q^{m-k},\ ae^{i\phi}q^m,\ ae^{-i\phi}q^m\\ bcq^m,\ a^2/bcq^{k-2m-1}\end{matrix};q\right]$$

$$=\frac{\left|\left(bce^{i\phi}/a\right)_k\right|^2}{(bc)_k\left(bc/a^2\right)_k}\,{}_{10}\phi_9\left[\begin{matrix}a^2/bcq^k,\ q\sqrt{\cdot},\ -q\sqrt{\cdot},\ a/b^2cq^{k-1},\ a/bc^2q^{k-1},\\ \sqrt{\cdot},\ -\sqrt{\cdot},\ ab,\ ac,\end{matrix}\right.$$

$$\left.\begin{matrix}ae^{i\theta},\ ae^{-i\theta},\ ae^{i\phi},\ ae^{-i\phi},\ q^{-k}\\ ae^{-i\theta}/bcq^{k-1},\ ae^{i\theta}/bcq^{k-1},\ ae^{-i\phi}/bcq^{k-1},\ ae^{i\phi}/bcq^{k-1},\ a^2q/bc\end{matrix};q\right]$$

by (3.23) and (3.12).

Hence, the first term on the r.h.s. of (5.5) equals

(5.10)

$$(1-t^2)\frac{(bcqt)_\infty}{(t/bc)_\infty}\sum_{k=0}^{\infty}\frac{(-bc)_k(bc\sqrt{q})_k(-bc\sqrt{q})_k|(bce^{i\theta}/a)_k(bce^{i\phi}/a)_k|^2q^k}{(q)_k(bc^2/a)_k(b^2c/a)_k(bc)_k(bc/a^2)_k(bcqt)_k(bcq/t)_k}$$

$$\cdot{}_{10}\phi_9\left[\begin{matrix}a^2/bcq^k, q\sqrt{\cdot}, -q\sqrt{\cdot}, a/b^2cq^{k-1}, a/bc^2q^{k-1}, ae^{i\theta},\\ \sqrt{\cdot}, -\sqrt{\cdot}, ab, ac, ae^{-i\theta}/bcq^{k-1},\end{matrix}\right.$$

$$\left.\begin{matrix}ae^{-i\theta}, ae^{i\phi}, ae^{-i\phi}, q^{-k}\\ ae^{i\theta}/bcq^{k-1}, ae^{-i\phi}/bcq^{k-1}, ae^{i\phi}/bcq^{k-1}, a^2q/bc\end{matrix};q\right].$$

Let us now compute the second term on the r.h.s. of (5.5). From (3.15),

(5.11)

$$\int_{qe^{i\phi}/c}^{qe^{-i\phi}/c}d_qv\frac{(bcvtq^{k-1})_\infty(cve^{i\phi})_\infty(cve^{-i\phi})_\infty(b^2c^2uv/q^2)_\infty}{(acv/q)_\infty(bcv/q)_\infty(bc^2v/aq)_\infty(bcuvtq^{k-2})_\infty}$$

$$=\frac{q(1-q)}{2ci}(q)_\infty(ab)_\infty(bc)_\infty(b^2c/a)_\infty|(ce^{i\phi})_\infty|^2w(y;a,b,c,bc/a;q)$$

$$\cdot\frac{(btq^ke^{-i\phi})_\infty(b^2cuq^{-1}e^{-i\phi})_\infty}{(butq^{k-1}e^{-i\phi})_\infty(b^2ce^{-i\phi})_\infty}$$

$$\cdot{}_8\phi_7\left[\begin{matrix}b^2ce^{-i\phi}/q, q\sqrt{\cdot}, -q\sqrt{\cdot}, bce^{-i\phi}/a, be^{-i\phi}, q/u, bct^{-1}q^{-k}, ae^{-i\phi}\\ \sqrt{\cdot}, -\sqrt{\cdot}, ab, bc, b^2cue^{-i\phi}/q, btq^ke^{-i\phi}, b^2c/a\end{matrix};butq^{k-1}e^{i\phi}\right].$$

Unfortunately, since this ${}_8\phi_7$ does not terminate we cannot apply formula (3.23) to it as in (5.6). However, we can still apply (3.15) to it to obtain

$$\int_{qe^{i\phi}/c}^{qe^{-i\phi}/c}d_qv\ldots$$

$$=\frac{q(1-q)(q)_\infty(bc)_\infty(b^2cu/aq)_\infty(abutq^k)_\infty(tbq^k/a)_\infty|(be^{i\phi})_\infty(ce^{i\phi})_\infty|^2}{2ci(b/a)_\infty(butq^{k-1}e^{i\phi})_\infty(butq^{k-1}e^{-i\phi})_\infty}$$

$$\cdot w(y;a,b,c,bc/a;q)\,{}_4\phi_3\left[\begin{matrix}tq^k, bcu/q, ae^{i\phi}, ae^{-i\phi}\\ bc, aq/b, abutq^{k-1}\end{matrix};q\right]$$

$$+\frac{q(1-q)(q)_\infty(bcu/q)_\infty(b^2c/a)_\infty(tq^k)_\infty(b^2utq^{k-1})_\infty|(ae^{i\phi})_\infty(ce^{i\phi})_\infty|^2}{2ci(a/b)_\infty(butq^{k-1}e^{i\phi})_\infty(butq^{k-1}e^{-i\phi})_\infty}$$

$$\cdot w(y;a,b,c,bc/a;q)\,{}_4\phi_3\left[\begin{matrix}tbq^k/a, b^2cu/aq, be^{i\phi}, be^{-i\phi}\\ bq/a, b^2c/a, b^2utq^{k-1}\end{matrix};q\right].$$

Therefore,

(5.12)

$$\int_{qe^{i\theta}/b}^{qe^{-i\theta}/b} d_q u \frac{(bcutq^{k-1})_\infty (bue^{i\theta})_\infty (bue^{-i\theta})_\infty}{(abu/q)_\infty (bcu/q)_\infty (b^2cu/aq)_\infty} \int_{qe^{i\theta}/c}^{qe^{-i\theta}/c} d_q v \dots$$

$$= \frac{q(1-q)(q)_\infty}{2ci(b/a)_\infty} (bc)_\infty (btq^k/a)_\infty \left|(be^{i\phi})_\infty (ce^{i\phi})_\infty\right|^2 w(y;\, a,b,c,bc/a;\, q)$$

$$\cdot \sum_{r=0}^{\infty} \frac{(tq^k)_r \left|(ae^{i\phi})_r\right|^2 q^r}{(q)_r (bc)_r (aq/b)_r}$$

$$\cdot \int_{qe^{i\theta}/b}^{qe^{-i\theta}/b} d_q u \frac{(bcutq^{k-1})_\infty (bue^{i\theta})_\infty (bue^{-i\theta})_\infty (abutq^{k+r-1})_\infty}{(abu/q)_\infty (bcuq^{r-1})_\infty (butq^{k-1}e^{i\phi})_\infty (butq^{k-1}e^{-i\phi})_\infty}$$

$$+ \frac{q(1-q)(q)_\infty}{2ci(a/b)_\infty} (b^2c/a)_\infty (tq^k)_\infty \left|(ae^{i\phi})_\infty (ce^{i\phi})_\infty\right|^2 w(y;\, a,b,c,bc/a;\, q)$$

$$\cdot \sum_{r=0}^{\infty} \frac{(btq^k/a)_r \left|(be^{i\phi})_r\right|^2 q^r}{(q)_r (bq/a)_r (b^2c/a)_r}$$

$$\cdot \int_{qe^{i\theta}/b}^{qe^{-i\theta}/b} \frac{(bcutq^{k-1})_\infty (bue^{i\theta})_\infty (bue^{-i\theta})_\infty (b^2utq^{k+r-1})_\infty}{(abu/q)_\infty (butq^{k-1}e^{i\phi})_\infty (butq^{k-1}e^{-i\phi})_\infty (b^2cuq^{r-1}/a)_\infty}.$$

Since, by (3.15) and (3.19),

$$\int_{qe^{i\theta}/b}^{qe^{-i\theta}/b} d_q u \frac{(bcutq^{k-1})_\infty (bue^{i\theta})_\infty (bue^{-i\theta})_\infty (abutq^{k+r-1})_\infty}{(abu/q)_\infty (bcuq^{r-1})_\infty (butq^{k-1}e^{i\phi})_\infty (butq^{k-1}e^{-i\phi})_\infty}$$

$$= \frac{q(1-q)}{2bi\sin\theta} \frac{(q)_\infty (acq^r)_\infty \left|(e^{2i\theta})_\infty\right|^2 (atq^k e^{i\phi})_\infty (ctq^{k+r}e^{i\phi})_\infty}{(actq^{k+r}e^{i\phi-i\theta})_\infty \left|(tq^k e^{i\theta+i\phi})_\infty (ae^{i\theta})_\infty (cq^r e^{i\theta})_\infty\right|^2}$$

$$\cdot \frac{(ctq^k e^{-i\theta})_\infty (atq^{k+r}e^{-i\theta})_\infty}{(tq^k e^{i\phi-i\theta})_\infty} {}_8\Phi_7\left[\begin{matrix} actq^{k+r-1}e^{i\phi-i\theta},\ q\sqrt{\cdot},\ -q\sqrt{\cdot},\ tq^k e^{i\phi-i\theta}, \\ \sqrt{\cdot},\ -\sqrt{\cdot},\ acq^r, \\ aq^r e^{i\phi},\ cq^r e^{-i\theta},\ ce^{i\phi},\ ae^{-i\theta} \\ ctq^k e^{-i\theta},\ atq^k e^{i\phi},\ atq^{k+r}e^{-i\theta},\ ctq^{k+r}e^{i\phi} \end{matrix};\, tq^k e^{i\theta-i\phi}\right]$$

$$= \frac{q(1-q)}{2bi} w(x;\, a,b,c,bc/a;\, q)(q)_\infty (ac)_\infty (tq^k)_\infty \left|(be^{i\theta})_\infty (bca^{-1}e^{i\theta})_\infty\right|^2$$

$$\cdot \frac{\left|(atq^k e^{i\phi})_\infty (ctq^k e^{i\theta})_\infty (ce^{i\theta})_r\right|^2}{\left|(tq^k e^{i\theta+i\phi})_\infty (tq^k e^{i\theta-i\phi})_\infty\right|^2 (actq^k)_\infty (ac)_r (tq^k)_r}$$

$$\cdot {}_8\Phi_7\left[\begin{matrix} actq^{k-1},\ q\sqrt{\cdot},\ -q\sqrt{\cdot},\ tq^{k-r},\ ae^{i\theta},\ ae^{-i\theta},\ ce^{i\phi},\ ce^{-i\phi} \\ \sqrt{\cdot},\ -\sqrt{\cdot},\ acq^r,\ ctq^k e^{-i\theta},\ ctq^k e^{i\theta},\ atq^k e^{-i\phi},\ atq^k e^{i\phi} \end{matrix};\, tq^{k+r}\right],$$

the first term on the r.h.s. of (5.12) equals

(5.13)

$$-\frac{[q(1-q)(q)_\infty]^2}{4bc}w(x;\,a,b,c,bc/a;\,q)w(y;\,a,b,c,bc/a;\,q)_\infty(ac)_\infty(bc)_\infty$$

$$\cdot\frac{(tq^k)_\infty(btq^k/a)_\infty\left|(be^{i\phi})_\infty(ce^{i\phi})_\infty(be^{i\theta})_\infty(bce^{i\theta}/a)_\infty(atq^ke^{i\phi})_\infty(ctq^ke^{i\theta})_\infty\right|^2}{(b/a)_\infty(actq^k)_\infty\left|(tq^ke^{i\theta+i\phi})_\infty(tq^ke^{i\theta-i\phi})_\infty\right|^2}$$

$$\cdot\sum_{r=0}^{\infty}\frac{\left|(ae^{i\phi})_r(ce^{i\theta})_r\right|^2q^r}{(q)_r(ac)_r(bc)_r(aq/b)_r}$$

$$\cdot{}_8\phi_7\left[\begin{matrix}actq^{k-1},\ q\sqrt{\cdot},\ -q\sqrt{\cdot},\ tq^{k-r},\ ae^{i\theta},\ ae^{-i\theta},\ ce^{i\phi},\ ce^{-i\phi}\\ \sqrt{\cdot},\ -\sqrt{\cdot},\ acq^r,\ ctq^ke^{-i\theta},\ ctq^ke^{i\theta},\ atq^ke^{-i\phi},\ atq^ke^{i\phi}\end{matrix};\,tq^{k+r}\right].$$

Similarly, since

$$\int_{qe^{i\theta}/b}^{qe^{-i\theta}/b}d_qu\,\frac{(bcutq^{k-1})_\infty(bue^{i\theta})_\infty(bue^{-i\theta})_\infty(b^2utq^{k+r-1})_\infty}{(abu/q)_\infty(butq^{k-1}e^{i\phi})_\infty(butq^{k-1}e^{-i\phi})_\infty(b^2cuq^{r-1}/a)_\infty}$$

$$=\frac{q(1-q)}{2bi\sin\theta}\,\frac{(q)_\infty(bcq^r)_\infty\left|(e^{2i\theta})_\infty\right|^2(btq^{k+r}e^{-i\theta})_\infty}{(bctq^{k+r}e^{i\phi-i\theta})_\infty(tq^ke^{i\phi-i\theta})_\infty}$$

$$\cdot\frac{(ctq^ke^{-i\theta})_\infty(atq^ke^{i\phi})_\infty(bca^{-1}tq^{k+r}e^{i\phi})_\infty}{\left|(ae^{i\theta})_\infty(bca^{-1}q^re^{i\theta})_\infty(tq^ke^{i\theta+i\phi})_\infty\right|^2}$$

$$\cdot{}_8\phi_7\left[\begin{matrix}bctq^{k+r-1}e^{i\phi-i\theta},\ q\sqrt{\cdot},\ -q\sqrt{\cdot},\ bq^re^{i\phi},\ bca^{-1}q^re^{-i\theta},\ tq^ke^{i\phi-i\theta},\\ \sqrt{\cdot},\ -\sqrt{\cdot},\ ctq^ke^{-i\theta},\ atq^ke^{i\phi},\ bcq^r,\\ \qquad\qquad ae^{-i\theta},\ ce^{i\phi}\\ bca^{-1}tq^{k+r}e^{i\phi},\ btq^{k+r}e^{-i\theta}\end{matrix};\,tq^ke^{i\theta-i\phi}\right]$$

$$=\frac{q(1-q)}{2bi\sin\theta}\,\frac{(q)_\infty(bcq^r)_\infty(ba^{-1}tq^{k+r})_\infty\left|(e^{2i\theta})_\infty(atq^ke^{i\phi})_\infty(ctq^ke^{i\theta})_\infty\right|^2}{(actq^k)_\infty\left|(ae^{i\theta})_\infty(bca^{-1}q^re^{i\theta})_\infty(tq^ke^{i\theta+i\phi})_\infty(tq^ke^{i\theta-i\phi})_\infty\right|^2}$$

$$\cdot{}_8\phi_7\left[\begin{matrix}actq^{k-1},\ q\sqrt{\cdot},\ -q\sqrt{\cdot},\ ab^{-1}tq^{k-r},\ ae^{i\theta},\ ae^{-i\theta},\ ce^{i\phi},\ ce^{-i\phi}\\ \sqrt{\cdot},\ -\sqrt{\cdot},\ bcq^r,\ ctq^ke^{-i\theta},\ ctq^ke^{i\theta},\ atq^ke^{-i\phi},\ atq^ke^{i\phi}\end{matrix};\,ba^{-1}tq^{k+r}\right],$$

the second term on the r.h.s. of (5.12) equals

(5.14)

$$-\frac{[q(1-q)(q)_\infty]^2}{4bc}w(x;\,a,b,c,bc/a;\,q)w(y;\,a,b,c,bc/a;\,q)(bc)_\infty(b^2c/a)_\infty$$

$$\cdot\frac{(tq^k)_\infty(ba^{-1}tq^k)_\infty\left|(ae^{i\phi})_\infty(be^{i\theta})_\infty(ce^{i\theta})_\infty(ce^{i\phi})_\infty(atq^ke^{i\phi})_\infty(ctq^ke^{i\theta})_\infty\right|^2}{(a/b)_\infty(actq^k)_\infty\left|(tq^ke^{i\theta+i\phi})_\infty(tq^ke^{i\theta-i\phi})_\infty\right|^2}$$

$$\cdot\sum_{r=0}^{\infty}\frac{\left|(be^{i\phi})_r(bca^{-1}e^{i\theta})_r\right|^2q^r}{(q)_r(bc)_r(bq/a)_r(b^2c/a)_r}$$

$$\cdot{}_8\phi_7\left[\begin{matrix}actq^{k-1},\ q\sqrt{\cdot},\ -q\sqrt{\cdot},\ ab^{-1}tq^{k-r},\ ae^{i\theta},\ ae^{-i\theta},\ ce^{i\phi},\ ce^{-i\phi}\\ \sqrt{\cdot},\ -\sqrt{\cdot},\ bcq^r,\ ctq^ke^{-i\theta},\ ctq^ke^{i\theta},\ atq^ke^{-i\phi},\ atq^ke^{i\phi}\end{matrix};ba^{-1}tq^{k+r}\right].$$

At this stage it is possible to apply (3.19) to the ${}_8\phi_7$ series in (5.13) and (5.14) to get a representation for $K_t(x,y;\,a,b,c,bc/a;\,q)$ as the sum (5.10) plus two double sums of ${}_8\phi_7$ series which shows that this kernel is nonnegative in the continuous q-Jacobi case (4.9) for $\alpha,\beta>-1$ when $0\leqq t<q^{1/2}$. But in order to handle the full range $0\leqq t<1$ we need to use the deeper procedure in the next section.

6. Positivity of $P_t(x,y;\,a,\,b,\,c,\,bc/a;\,q)$. The above-mentioned procedure involves first using (3.15) to see that the ${}_8\phi_7$ series in (5.13) equals

$$\frac{(actq^k)_\infty(atq^k/c)_\infty\left|(aq^re^{i\phi})_\infty\right|^2}{(acq^r)_\infty(aq^r/c)_\infty\left|(atq^ke^{i\phi})_\infty\right|^2}\,{}_4\phi_3\left[\begin{matrix}ctq^k/a,\ tq^{k-r},\ ce^{i\phi},\ ce^{-i\phi}\\ c/aq^{r-1},\ ctq^ke^{i\theta},\ ctq^ke^{-i\theta}\end{matrix};q\right]$$

$$+\frac{(actq^k)_\infty(ctq^k/a)_\infty(tq^{k-r})_\infty\left|(ce^{i\phi})_\infty(atq^{k+r}e^{i\theta})_\infty\right|^2}{(acq^r)_\infty(tq^{k+r})_\infty(c/aq^r)_\infty\left|(ctq^ke^{i\theta})_\infty(atq^ke^{i\phi})_\infty\right|^2}$$

$$\cdot{}_4\phi_3\left[\begin{matrix}atq^k/c,\ tq^{k+r},\ aq^re^{i\phi},\ aq^re^{-i\phi}\\ ac^{-1}q^{r+1},\ atq^{k+r}e^{i\theta},\ atq^{k+r}e^{-i\theta}\end{matrix};q\right]$$

and the ${}_8\phi_7$ series in (5.14) equals

$$\frac{(actq^k)_\infty(atq^k/c)_\infty\left|(bq^re^{i\phi})_\infty\right|^2}{(bcq^r)_\infty(bq^r/c)_\infty\left|(atq^ke^{i\phi})_\infty\right|^2}\,{}_4\phi_3\left[\begin{matrix}ctq^k/a,\ ab^{-1}tq^{k-r},\ ce^{i\phi},\ ce^{-i\phi}\\ c/bq^{r-1},\ ctq^ke^{i\theta},\ ctq^ke^{-i\theta}\end{matrix};q\right]$$

$$+\frac{(actq^k)_\infty(ctq^k/a)_\infty(ab^{-1}tq^{k-r})_\infty\left|(ce^{i\phi})_\infty(btq^{k+r}e^{i\theta})_\infty\right|^2}{(bcq^r)_\infty(ba^{-1}tq^{k+r})_\infty(c/bq^r)_\infty\left|(ctq^ke^{i\theta})_\infty(atq^ke^{i\phi})_\infty\right|^2}$$

$$\cdot{}_4\phi_3\left[\begin{matrix}atq^k/c,\ ba^{-1}tq^{k+r},\ bq^re^{i\phi},\ bq^re^{-i\phi},\\ bc^{-1}q^{r+1},\ btq^{k+r}e^{i\theta},\ btq^{k+r}e^{-i\theta}\end{matrix};q\right].$$

Using the above in (5.13) and (5.14) and writing

$$V(x,y) \equiv -\frac{[q(1-q)(q)_\infty]^2}{4bc} w(x;\, a,b,c,bc/a;\, q)w(y;\, a,b,c,bc/a;\, q)$$

it follows by regrouping that we can write the r.h.s. of (5.12) in the form $G_1(x,y)+G_2(x,y)$ with

(6.1)

$$G_1(x,y) = \frac{V(x,y)(b^2c/a)_\infty |(ae^{i\phi})_\infty (be^{i\phi})_\infty (ce^{i\phi})_\infty (be^{i\theta})_\infty (ce^{i\theta})_\infty (ctq^k e^{i\theta})_\infty|^2}{(b/c)_\infty (a/b)_\infty |(tq^k e^{i\theta+i\phi})_\infty (tq^k e^{i\theta-i\phi})_\infty|^2}$$

$$\cdot (btq^k/a)_\infty (atq^k/c)_\infty (tq^k)_\infty \sum_{r=0}^{\infty} \frac{(b/c)_r |(bce^{i\theta}/a)_r|^2 q^r}{(q)_r (bq/a)_r (b^2c/a)_r}$$

$$\cdot {}_4\phi_3\left[\begin{matrix} ctq^k/a,\ ab^{-1}tq^{k-r},\ ce^{i\phi},\ ce^{-i\phi} \\ c/bq^{r-1},\ ctq^k e^{i\theta},\ ctq^k e^{-i\theta} \end{matrix};\, q\right]$$

$$+ \frac{V(x,y)(bc)_\infty |(ae^{i\phi})_\infty (be^{i\phi})_\infty (ce^{i\phi})_\infty (be^{i\theta})_\infty (bce^{i\theta}/a)_\infty (ctq^k e^{i\theta})_\infty|^2}{(a/c)_\infty (b/a)_\infty |(tq^k e^{i\theta+i\phi})_\infty (tq^k e^{i\theta-i\phi})_\infty|^2}$$

$$\cdot (btq^k/a)_\infty (atq^k/c)_\infty (tq^k)_\infty \sum_{r=0}^{\infty} \frac{(a/c)_r |(ce^{i\theta})_r|^2 q^r}{(q)_r (bc)_r (aq/b)_r}$$

$$\cdot {}_4\phi_3\left[\begin{matrix} ctq^k/a,\ tq^{k-r},\ ce^{i\phi},\ ce^{-i\phi} \\ c/aq^{r-1},\ ctq^k e^{i\theta},\ ctq^k e^{-i\theta} \end{matrix};\, q\right]$$

and

(6.2)

$$G_2(x,y) = \frac{V(x,y)(b^2c/a)_\infty |(ae^{i\phi})_\infty (ce^{i\phi})_\infty (ce^{i\phi})_\infty (be^{i\theta})_\infty (ce^{i\theta})_\infty (btq^k e^{i\theta})_\infty|^2}{(c/b)_\infty (a/b)_\infty |(tq^k e^{i\theta+i\phi})_\infty (tq^k e^{i\theta-i\phi})_\infty|^2}$$

$$\cdot (ctq^k/a)_\infty (atq^k/b)_\infty (tq^k)_\infty \sum_{r=0}^{\infty}\sum_{s=0}^{\infty} \frac{(btq^k/a)_{r+s} |(be^{i\phi})_{r+s}|^2}{(q)_r (q)_s |(btq^k e^{i\theta})_{r+s}|^2}$$

$$\cdot \frac{(atq^k/c)_s |(bce^{i\theta}/a)_r|^2 (b/atq^{k-1})_r}{(bq/c)_{r+s} (bq/a)_r (b^2c/a)_r} (ac^{-1}tq^{k+1})^r q^s$$

$$+ \frac{V(x,y)(bc)_\infty |(be^{i\phi})_\infty (ce^{i\phi})_\infty (ce^{i\phi})_\infty (be^{i\theta})_\infty (bce^{i\theta}/a)_\infty (atq^k e^{i\theta})_\infty|^2}{(b/a)_\infty (c/a)_\infty |(tq^k e^{i\theta+i\phi})_\infty (tq^k e^{i\theta-i\phi})_\infty|^2}$$

$$\cdot (btq^k/a)_\infty (ctq^k/a)_\infty (tq^k)_\infty \sum_{r=0}^{\infty}\sum_{s=0}^{\infty} \frac{(tq^k)_{r+s} |(ae^{i\phi})_{r+s}|^2}{(q)_r (q)_s |(atq^k e^{i\theta})_{r+s}|^2}$$

$$\cdot \frac{(atq^k/c)_s (t^{-1}q^{1-k})_r |(ce^{i\theta})_r|^2}{(aq/c)_{r+s} (aq/b)_r (bc)_r} (ac^{-1}tq^{k+1})^r q^s.$$

By changing the orders of summation in both parts of $G_1(x,y)$ and using (3.15) it follows that

(6.3)

$$G_1(x,y)=\frac{V(x,y)(bc^2/a)_\infty\left|(ae^{i\phi})_\infty(be^{i\phi})_\infty(ce^{i\phi})_\infty(be^{i\theta})_\infty(be^{i\theta})_\infty(ctq^ke^{i\theta})_\infty\right|^2}{(a/c)_\infty(b/c)_\infty\left|(tq^ke^{i\theta+i\phi})_\infty(tq^ke^{i\theta-i\phi})_\infty\right|^2}$$

$$\cdot(btq^k/a)_\infty(atq^k/c)_\infty(tq^k)_\infty\sum_{j=0}^{\infty}\frac{(ctq^k/a)_j\left|(ce^{i\phi})_j(qb^{-1}e^{i\theta})_j\right|^2}{(q)_j(cq/b)_j(aq/b^2c)_j\left|(ctq^ke^{i\theta})_j\right|^2}$$

$$\cdot\frac{(atq^k/b)_j}{(cq/a)_j}q^k\,{}_8\phi_7\left[\begin{matrix}b^2c/aq^{j+1},\ q\sqrt{\cdot},\ -q\sqrt{\cdot},\ bctq^{k-1},\ bca^{-1}e^{i\theta},\\ \sqrt{\cdot},\ -\sqrt{\cdot},\ b/atq^{k+j-1},\ be^{-i\theta}q^{-j},\end{matrix}\right.$$

$$\left.\begin{matrix}bca^{-1}e^{-i\theta},\ b/cq^j,\ q^{-j}\\ be^{i\theta}q^{-j},\ bc^2/a,\ b^2c/a\end{matrix};t^{-1}q^{1-k}\right].$$

By using (3.23) to write the above ${}_8\phi_7$ as a multiple of the series

$${}_4\phi_3\left[\begin{matrix}q^{-j},\ c/atq^{k-1},\ bca^{-1}e^{i\theta},\ bca^{-1}e^{-i\theta}\\ bc^2/a,\ cq/a,\ b/atq^{k+j-1}\end{matrix};q\right]$$

and then using it again to write this ${}_4\phi_3$ as a multiple of the series

$${}_8\phi_7\left[\begin{matrix}bc^2a^{-1}tq^{k-1},\ q\sqrt{\cdot},\ -q\sqrt{\cdot},\ tq^k,\ bctq^{k-1},\ bca^{-1}e^{i\theta},\\ \sqrt{\cdot},\ -\sqrt{\cdot},\ bc^2/a,\ cq/a,\ ctq^ke^{-i\theta},\end{matrix}\right.$$

$$\left.\begin{matrix}bca^{-1}e^{-i\theta},\ q^{-j}\\ ctq^ke^{i\theta},\ bc^2a^{-1}tq^{k+j}\end{matrix};cb^{-1}q^{j+1}\right]$$

we find that the sum over j in (6.3) equals

(6.4)

$$\sum_{j=0}^{\infty}\frac{(ctq^k/a)_j\left|(ce^{i\phi})_j\right|^2}{(q)_j(cq/b)_j(bc^2tq^k/a)_j}q^j\sum_{r=0}^{j}\frac{(bc^2a^{-1}tq^{k-1})_r(q\sqrt{\cdot})_r(-q\sqrt{\cdot})_r}{(q)_r(\sqrt{\cdot})_r(-\sqrt{\cdot})_r}$$

$$\cdot\frac{(tq^k)_r(bctq^{k-1})_r\left|(bca^{-1}e^{i\theta})_r\right|^2(q)_j(-1)^r(cq/b)^r}{(bc^2/a)_r(cq/a)_r\left|(ctq^ke^{i\theta})_r\right|^2(bc^2a^{-1}tq^{k+j})_r(q)_{j-r}}q^{r(r-1)/2}$$

$$=\sum_{r=0}^{\infty}\frac{(bc^2a^{-1}tq^{k-1})_r(q\sqrt{\cdot})_r(-q\sqrt{\cdot})_r(tq^k)_r(bctq^{k-1})_r(ctq^k/a)_r\left|(ce^{i\phi})_r\right|^2}{(q)_r(\sqrt{\cdot})_r(-\sqrt{\cdot})_r(bc^2/a)_r(cq/a)_r(cq/b)_r\left|(ctq^re^{i\theta})_r\right|^2}$$

$$\cdot\frac{\left|(bca^{-1}e^{i\theta})_r\right|^2}{(bc^2tq^k/a)_{2r}}(-1)^r\left(\frac{cq}{b}\right)^r q^{r(r+1)/2}\,{}_3\phi_2\left[\begin{matrix}ca^{-1}tq^{k+r},\ cq^re^{i\phi},\ cq^re^{-i\phi}\\ cb^{-1}q^{r+1},\ bc^2a^{-1}tq^{k+2r}\end{matrix};q\right].$$

Fortunately, we can now apply (3.10) to the above ${}_3\phi_2$ to show that it equals

$$\frac{(b/cq^r)_\infty (bcq^r)_\infty \left|(bca^{-1}tq^{k+r}e^{i\phi})_\infty\right|^2}{(btq^k/a)_\infty (bc^2a^{-1}tq^{k+2r})_\infty \left|(be^{i\phi})_\infty\right|^2}$$
$$+ bc^{-1}q^{-r}\frac{(ca^{-1}tq^{k+r})_\infty (b/cq^{r-1})_\infty (b^2ca^{-1}tq^{k+r})_\infty \left|(cq^re^{i\phi})_\infty\right|^2}{(cb^{-1}q^{r+1})_\infty (bc^2a^{-1}tq^{k+2r})_\infty (btq^k/a)_\infty \left|(be^{i\phi})_\infty\right|^2}$$
$$\cdot {}_3\phi_2\left[\begin{matrix} btq^k/a,\ be^{i\phi},\ be^{-i\phi} \\ b/cq^{r-1},\ b^2ca^{-1}tq^{k+r}\end{matrix};q\right].$$

Using this in (6.4) we obtain that the sum over j in (6.3) equals

(6.5)

$$\frac{(b/c)_\infty (bc)_\infty \left|(bca^{-1}tq^ke^{i\phi})_\infty\right|^2}{(btq^k/a)_\infty (bc^2tq^k/a)_\infty \left|(be^{i\phi})_\infty\right|^2}$$
$$\cdot {}_{10}\phi_9\left[\begin{matrix} bc^2a^{-1}tq^{k-1},\ q\sqrt{\cdot},\ -q\sqrt{\cdot},\ tq^k, \\ \sqrt{\cdot},\ -\sqrt{\cdot},\ bc^2/a, \\ bctq^{k-1},\ bca^{-1}e^{i\theta},\ bca^{-1}e^{-i\theta},\ ce^{i\phi},\ ce^{-i\phi},\ ctq^k/a \\ cq/a,\ ctq^ke^{-i\theta},\ ctq^ke^{i\theta},\ bca^{-1}tq^ke^{-i\phi},\ bca^{-1}tq^ke^{i\phi},\ bc\end{matrix};q\right]$$
$$- \frac{(b/c)_\infty (ctq^k/a)_\infty (b^2ctq^k/a)_\infty \left|(ce^{i\phi})_\infty\right|^2}{(c/b)_\infty (bc^2tq^k/a)_\infty (btq^k/a)_\infty \left|(be^{i\phi})_\infty\right|^2}$$
$$\cdot \sum_{j=0}^{\infty}\frac{(btq^k/a)_j \left|(be^{i\phi})_j\right|^2}{(q)_j (bq/c)_j (b^2ctq^k/a)_j}q^j$$
$$\cdot {}_8\phi_7\left[\begin{matrix} bc^2a^{-1}tq^{k-1},\ q\sqrt{\cdot},\ -q\sqrt{\cdot},\ tq^k,\ bctq^{k-1},\ bca^{-1}e^{i\theta},\ bca^{-1}e^{-i\theta},\ c/bq^j \\ \sqrt{\cdot},\ -\sqrt{\cdot},\ bc^2/a,\ cq/a,\ ctq^ke^{-i\theta},\ ctq^ke^{i\theta},\ b^2ca^{-1}tq^{k+j}\end{matrix};q^{j+1}\right]$$

by a change in order of summation.

Let us now consider $G_2(x,y)$. From (6.2)

(6.6)

$$G_2(x,y) = \frac{V(x,y)(b^2c/a)_\infty \left|(ae^{i\phi})_\infty (ce^{i\phi})_\infty (ce^{i\phi})_\infty (be^{i\theta})_\infty (ce^{i\theta})_\infty (btq^ke^{i\theta})\right|^2}{(c/b)_\infty (a/b)_\infty \left|(tq^ke^{i\theta+i\phi})_\infty (tq^ke^{i\theta-i\phi})_\infty\right|^2}$$
$$\cdot (ctq^k/a)_\infty (atq^k/b)_\infty (tq^k)_\infty$$
$$\cdot \sum_{j=0}^{\infty}\frac{(btq^k/a)_j (atq^k/c)_j \left|(be^{i\phi})_j\right|^2}{(q)_j (bq/c)_j \left|(btq^ke^{i\theta})_j\right|^2}q^j$$
$$\cdot {}_4\phi_3\left[\begin{matrix} q^{-j},\ b/atq^{k-1},\ bca^{-1}e^{i\theta},\ bca^{-1}e^{-i\theta} \\ bq/a,\ b^2c/a,\ c/atq^{k+j-1}\end{matrix};q\right]$$

$$+\frac{V(x,y)(bc)_\infty\left|(be^{i\phi})_\infty(ce^{i\phi})_\infty(ce^{i\phi})_\infty(be^{i\theta})_\infty(bca^{-1}e^{i\theta})_\infty(atq^ke^{i\theta})_\infty\right|^2}{(b/a)_\infty(c/a)_\infty\left|(tq^ke^{i\theta+i\phi})_\infty(tq^ke^{i\theta-i\phi})_\infty\right|^2}$$

$$\cdot(btq^k/a)_\infty(ctq^k/a)_\infty(tq^k)_\infty\sum_{j=0}^{\infty}\frac{(tq^k)_j(atq^k/c)_j\left|(ae^{i\phi})_j\right|^2}{(q)_j(aq/c)_j\left|(atq^ke^{i\theta})_j\right|^2}q^j$$

$$\cdot{}_4\phi_3\left[\begin{matrix}q^{-j},\ t^{-1}q^{1-k},\ ce^{i\theta},\ ce^{-i\theta}\\ bc,\ aq/b,\ c/atq^{k+j-1}\end{matrix};q\right].$$

Using (3.23), the first ${}_4\phi_3$ on the r.h.s. of (6.6) equals

(6.7)

$$\frac{\left|(btq^ke^{i\theta})_j\right|^2}{(atq^k/c)_j(b^2ctq^k/a)_j}{}_8\phi_7\left[\begin{matrix}b^2ca^{-1}tq^{k-1},\ q\sqrt{\cdot},\ -\sqrt{\cdot},\ bctq^{k-1},\ tq^k,\\ \sqrt{\cdot},\ -\sqrt{\cdot},\ bq/a,\ b^2c/a,\\ bca^{-1}e^{i\theta},\ bca^{-1}e^{-i\theta},\ q^{-j}\\ btq^ke^{-i\theta},\ btq^ke^{i\theta},\ b^2ca^{-1}tq^{k+j}\end{matrix};bc^{-1}q^{j+1}\right]$$

$$=\frac{(b^2ca^{-1}tq^{k+j})_\infty(ac^{-1}tq^{k+j})_\infty\left|(b^2a^{-1}q^{j+1}e^{i\theta})_\infty\right|^2}{(bc^{-1}q^{j+1})_\infty(b^3ca^{-2}q^{j+1})_\infty\left|(btq^{k+j}e^{i\theta})_\infty\right|^2}$$

$$\cdot{}_8\phi_7\left[\begin{matrix}b^3ca^{-2}q^j,\ q\sqrt{\cdot},\ -q\sqrt{\cdot},\ b^2cq^j/a,\ ba^{-1}q^{j+1},\ b/atq^{k-1},\\ \sqrt{\cdot},\ -\sqrt{\cdot},\ bq/a,\ b^2c/a,\ b^2ca^{-1}tq^{k+j},\\ bca^{-1}e^{i\theta},\ bca^{-1}e^{-i\theta}\\ b^2a^{-1}q^{j+1}e^{-i\theta},\ b^2a^{-1}q^{j+1}e^{i\theta}\end{matrix};ac^{-1}tq^k\right]$$

by (3.19). Our reason for choosing the above ${}_8\phi_7$ from among all the other possibilities will be clear from (6.12) and (6.11) below. Since the second ${}_4\phi_3$ on the r.h.s. of (6.6) equals

(6.8)
$$\frac{(aq/c)_j(b/atq^{k+j-1})_j}{(aq/b)_j(c/atq^{k+j-1})_j}\left(\frac{c}{b}\right)^j{}_4\phi_3\left[\begin{matrix}q^{-j},\ t^{-1}q^{1-k},\ be^{i\theta},\ be^{-i\theta}\\ bc,\ aq/c,\ b/atq^{k+j-1}\end{matrix};q\right]$$

$$=\frac{(aq/c)_j\left|(atq^ke^{i\theta})_j\right|^2}{(aq/b)_j(atq^k/c)_j(abtq^k)_j}$$

$$\cdot{}_8\phi_7\left[\begin{matrix}abtq^{k-1},\ q\sqrt{jdot},\ -q\sqrt{\cdot},\ atq^k/c,\ bctq^{k-1},\\ \sqrt{\cdot},\ -\sqrt{\cdot},\ bc,\ aq/c,\\ be^{i\theta},\ be^{-i\theta},\ q^{-j}\\ atq^ke^{-i\theta},\ atq^ke^{i\theta},\ abtq^{k+j}\end{matrix};ab^{-1}q^{j+1}\right],$$

the second sum over j on the r.h.s. of (6.6) equals

$$(6.9)\quad \sum_{r=0}^{\infty} \frac{(abtq^{k-1})_r(q\sqrt{\cdot})_r(-q\sqrt{\cdot})_r(atq^k/c)_r(bctq^{k-1})_r(tq^k)_r(-1)^r q^{r(r+1)/2}}{(q)_r(\sqrt{\cdot})_r(-\sqrt{\cdot})_r(bc)_r(aq/c)_r(aq/b)_r(abtq^k)_{2r}} \cdot \frac{|(ae^{i\phi})_r(be^{i\theta})_r|^2}{|(atq^k e^{i\theta})_r|^2}\left(\frac{aq}{b}\right)^r {}_3\phi_2\left[\begin{matrix} tq^{k+r}, aq^r e^{i\phi}, aq^r e^{-i\phi} \\ ab^{-1}q^{r+1}, abtq^{k+2r}\end{matrix}; q\right]$$

by a change in order of summation. We can apply (3.10) to the above ${}_3\phi_2$ to find that it equals

$$\frac{(b/aq^r)_\infty(abq^r)_\infty|(btq^{k+r}e^{i\phi})_\infty|^2}{(abtq^{k+2r})_\infty(btq^k/a)_\infty|(be^{i\phi})_\infty|^2} - \frac{(tq^{k+r})_\infty(b/aq^r)_\infty(b^2tq^{k+r})_\infty|(aq^r e^{i\phi})_\infty|^2}{(aq^r/b)_\infty(btq^k/a)_\infty(abtq^{k+2r})_\infty|(be^{i\phi})_\infty|^2} {}_3\phi_2\left[\begin{matrix} btq^k/a, be^{i\phi}, be^{-i\phi} \\ b/aq^{r-1}, b^2tq^{k+r}\end{matrix}; q\right].$$

Hence, from (6.9), the second sum over j on the r.h.s. of (6.6) equals

(6.10)

$$\frac{(ab)_\infty(b/a)_\infty|(btq^k e^{i\phi})_\infty|^2}{(abtq^k)_\infty(btq^k/a)_\infty|(be^{i\phi})_\infty|^2} {}_{10}\phi_9\left[\begin{matrix} abtq^{k-1}, q\sqrt{\cdot}, -q\sqrt{\cdot}, atq^k/c, \\ \sqrt{\cdot}, -\sqrt{\cdot}, bc, \\ bctq^{k-1}, tq^k, be^{i\theta}, be^{-i\theta}, ae^{i\phi}, ae^{-i\phi} \\ aq/c, ab, atq^k e^{-i\theta}, atq^k e^{i\theta}, btq^k e^{-i\phi}, btq^k e^{i\phi}\end{matrix}; q\right]$$

$$- \frac{(b/a)_\infty(tq^k)_\infty(b^2tq^k)_\infty|(ae^{i\phi})_\infty|^2}{(a/b)_\infty(abtq^k)_\infty(btq^k/a)_\infty|(be^{i\phi})_\infty|^2} \sum_{j=0}^{\infty} \frac{(btq^k/a)_j|(be^{i\phi})_j|^2}{(q)_j(bq/a)_j(b^2tq^k)_j} q^j \cdot {}_8\phi_7\left[\begin{matrix} abtq^{k-1}, q\sqrt{\cdot}, -q\sqrt{\cdot}, atq^k/c, bctq^{k-1}, be^{i\theta}, be^{-i\theta}, a/bq^j \\ \sqrt{\cdot}, -\sqrt{\cdot}, bc, aq/c, atq^k e^{-i\theta}, atq^k e^{i\theta}, b^2tq^{k+j}\end{matrix}; q^{j+1}\right]$$

by a change in order of summation.

Combining (6.3), (6.5), (6.6), (6.7) and (6.10) we obtain that

(6.11)

$$G_1(x,y)+G_2(x,y)=\frac{V(x,y)(tq^k)_\infty(bc)_\infty|(ce^{i\phi})_\infty(be^{i\theta})_\infty|^2}{|(tq^k e^{i\theta+i\phi})_\infty(tq^k e^{i\theta-i\phi})_\infty|^2} \cdot \left\{\frac{(bc^2/a)_\infty(atq^k/c)_\infty}{(a/c)_\infty(bc^2tq^k/a)_\infty}|(ae^{i\phi})_\infty(be^{i\theta})_\infty(ctq^k e^{i\theta})_\infty(bca^{-1}tq^k e^{i\phi})_\infty|^2 \cdot {}_{10}\phi_9\left[\begin{matrix} bc^2a^{-1}tq^{k-1}, q\sqrt{\cdot}, -q\sqrt{\cdot}, \\ \sqrt{\cdot}, -\sqrt{\cdot}, \\ tq^k, bctq^{k-1}, ctq^k/a, ce^{i\phi}, ce^{-i\phi}, bca^{-1}e^{i\theta}, bca^{-1}e^{-i\theta} \\ bc^2/a, cq/a, bc, bca^{-1}tq^k e^{-i\phi}, bca^{-1}tq^k e^{i\phi}, ctq^k e^{-i\theta}, ctq^k e^{i\theta}\end{matrix}; q\right]\right.$$

$$+\frac{(ab)_\infty(ctq^k/a)_\infty}{(c/a)_\infty(abtq^k)_\infty}\left|(ce^{i\phi})_\infty(bca^{-1}e^{i\theta})_\infty(atq^ke^{i\theta})_\infty(btq^ke^{i\phi})_\infty\right|^2$$

$$\cdot{}_{10}\phi_9\left[\begin{matrix}abtq^{k-1}, q\sqrt{\cdot}, -q\sqrt{\cdot}, atq^k/c, bctq^{k-1}, tq^k, be^{i\theta}, be^{-i\theta},\\ \sqrt{\cdot}, -\sqrt{\cdot}, bc, aq/c, ab, atq^ke^{-i\theta}, atq^ke^{i\theta},\\ ae^{i\phi}, ae^{-i\phi}\\ btq^ke^{-i\phi}, btq^ke^{i\phi}\end{matrix};q\right]\Bigg\}$$

$$+\frac{V(x,y)(tq^k)_\infty(ctq^k/a)_\infty\left|(ae^{i\phi})_\infty(ce^{i\phi})_\infty(ce^{i\phi})_\infty(be^{i\theta})_\infty\right|^2}{\left|(tq^ke^{i\theta+i\phi})_\infty(tq^ke^{i\theta-i\phi})_\infty\right|^2}$$

$$\cdot\Bigg\{\frac{(b^2c/a)_\infty(atq^k/c)_\infty(b^2ctq^k/a)_\infty(atq^k/b)_\infty}{(a/b)_\infty(c/b)_\infty(b/c)_\infty(b^3cq/a^2)_\infty}\left|(ce^{i\theta})_\infty(b^2a^{-1}qe^{i\theta})_\infty\right|^2$$

$$\cdot\sum_{j=0}^{\infty}\frac{(btq^k/a)_j(b^3cq/a^2)_j\left|(be^{i\phi})_j\right|^2}{(q)_j(b^2ctq^k/a)_j\left|(b^2a^{-1}qe^{i\theta})_j\right|^2}q^j$$

$$\cdot{}_8\phi_7\left[\begin{matrix}b^3ca^{-2}q^j, q\sqrt{\cdot}, -q\sqrt{\cdot},\\ \sqrt{\cdot}, -\sqrt{\cdot},\\ b^2cq^j/a, ba^{-1}q^{j+1}, b/atq^{k-1}, bca^{-1}e^{i\theta}, bca^{-1}e^{-i\theta}\\ bq/a, b^2c/a, b^2ca^{-1}tq^{k+j}, b^2a^{-1}q^{j+1}e^{-i\theta}, b^2a^{-1}q^{j+1}e^{i\theta}\end{matrix};ac^{-1}tq^k\right]$$

$$-\frac{(bc)_\infty(b^2tq^k)_\infty(tq^k)_\infty}{(a/b)_\infty(c/a)_\infty(abtq^k)_\infty}\left|(bca^{-1}e^{i\theta})_\infty(atq^ke^{i\theta})_\infty\right|^2$$

$$\cdot\sum_{j=0}^{\infty}\frac{(btq^k/a)_j}{(q)_j(bq/a)_j}\frac{\left|(be^{i\phi})_j\right|^2}{(b^2tq^k)_j}q^j$$

$$\cdot{}_8\phi_7\left[\begin{matrix}abtq^{k-1}, q\sqrt{\cdot}, -q\sqrt{\cdot}, atq^k/c, bctq^{k-1}, a/bq^j,\\ \sqrt{\cdot}, -\sqrt{\cdot}, bc, aq/c, b^2tq^{k+j},\\ be^{i\theta}, be^{-i\theta}\\ atq^ke^{-i\theta}, atq^ke^{i\theta}\end{matrix};q^{j+1}\right]$$

$$-\frac{(bc^2/a)_\infty(atq^k/c)_\infty(b^2ctq^k/a)_\infty}{(a/c)_\infty(c/b)_\infty(bc^2tq^k/a)_\infty}\left|(be^{i\theta})_\infty(ctq^ke^{i\theta})_\infty\right|^2$$

$$\cdot\sum_{k=0}^{\infty}\frac{(btq^k/a)_j}{(q)_j(bq/c)_j}\frac{\left|(be^{i\phi})_j\right|^2}{(b^2ctq^k/a)_j}q^j$$

$$\cdot{}_8\phi_7\left[\begin{matrix}bc^2a^{-1}tq^{k-1}, q\sqrt{\cdot}, -q\sqrt{\cdot}, bctq^{k-1}, c/bq^j,\\ \sqrt{\cdot}, -\sqrt{\cdot}, cq/a, b^2ca^{-1}tq^{k+j},\\ tq^k, bca^{-1}e^{i\theta}, bca^{-1}e^{-i\theta}\\ bc^2/a, ctq^ke^{-i\theta}, ctq^ke^{i\theta}\end{matrix};q^{j+1}\right]\Bigg\}.$$

Now, we need but observe that, by Bailey [6, (5.1)], the last ${}_8\phi_7$ in (6.11) equals

(6.12)

$$\frac{(a/c)_\infty (b^2c/a)_\infty (atq^k/b)_\infty (bc^2tq^k/a)_\infty \left|(ce^{i\theta})_\infty (b^2a^{-1}q^{j+1}e^{i\theta})_\infty\right|^2}{(a/b)_\infty (bc^2/a)_\infty (bc^{-1}q^{j+1})_\infty \left|(be^{i\theta})_\infty (ctq^ke^{i\theta})_\infty\right|^2}$$

$$\cdot {}_8\phi_7\left[\begin{matrix} b^3ca^{-2}q^j,\ q\sqrt{\cdot},\ -q\sqrt{\cdot},\ b^2cq^j/a,\ ba^{-1}q^{j+1},\ b/atq^{k-1} \\ \sqrt{\cdot},\ -\sqrt{\cdot},\ bq/a,\ b^2c/a,\ b^2ca^{-1}tq^{k+j}, \end{matrix}\right.$$

$$\left.\begin{matrix} bca^{-1}e^{i\theta},\ bca^{-1}e^{-i\theta} \\ b^2a^{-1}q^{j+1}e^{-i\theta},\ b^2a^{-1}q^{j+1}e^{i\theta} \end{matrix};\ ac^{-1}tq^k\right]$$

$$- \frac{(bc)_\infty (a/c)_\infty (tq^k)_\infty (ba^{-1}q^{j+1})_\infty (bc^2tq^k/a)_\infty (b^2tq^{k+j})_\infty}{(c/a)_\infty (bc^2/a)_\infty (abtq^k)_\infty (atq^k/c)_\infty (bc^{-1}q^{j+1})_\infty (b^2c^{-1}atq^{k+j})_\infty}$$

$$\cdot \frac{\left|(atq^ke^{i\theta})_\infty (bca^{-1}e^{i\theta})_\infty\right|^2}{\left|(be^{i\theta})_\infty (ctq^ke^{i\theta})_\infty\right|^2} \cdot \frac{(c/b)_\infty (bq/c)_\infty}{(a/b)_\infty (bq/a)_\infty}$$

$$\cdot {}_8\phi_7\left[\begin{matrix} abtq^{k-1},\ q\sqrt{\cdot},\ -q\sqrt{\cdot},\ atq^k/c,\ bctq^{k-1},\ a/bq^j,\ be^{i\theta},\ be^{-i\theta} \\ \sqrt{\cdot},\ -\sqrt{\cdot},\ bc,\ aq/c,\ b^2tq^{k+j},\ atq^ke^{-i\theta},\ atq^ke^{i\theta} \end{matrix};\ q^{j+1}\right],$$

from which it fortunately follows that the second expression in braces on the r.h.s. of (6.11) equals zero.

Hence, combining (5.5), (5.10), and (6.11) with the above observation we finally get our desired formula.

(6.13)

$$K_t(x,y;\ a,b,c,bc/a;\ q) = (1-t^2)\frac{(bcqt)_\infty}{(t/bc)_\infty}$$

$$\cdot \sum_{k=0}^{\infty} \frac{(-bc)_k}{(q)_k} \frac{(bc\sqrt{q})_k (-bc\sqrt{q})_k \left|(bca^{-1}e^{i\theta})_k (bca^{-1}e^{i\phi})_k\right|^2 q^k}{(bc^2/a)_k (b^2c/a)_k (bc)_k (bca^{-2})_k (bcqt)_k (bcq/t)_k}$$

$$\cdot {}_{10}\phi_9\left[\begin{matrix} a^2/bcq^k,\ q\sqrt{\cdot},\ -q\sqrt{\cdot},\ a/b^2cq^{k-1},\ a/bc^2q^{k-1},\ q^{-k}, \\ \sqrt{\cdot},\ -\sqrt{\cdot},\ ab,\ ac,\ a^2q/bc,\ ab^{-1}c^{-1}q^{1-k}e^{-i\theta}, \end{matrix}\right.$$

$$\left.\begin{matrix} ae^{i\theta},\ ae^{-i\theta},\ ae^{i\phi},\ ae^{-i\phi} \\ ab^{-1}c^{-1}q^{1-k}e^{i\theta},\ ab^{-1}c^{-1}q^{1-k}e^{-i\phi},\ ab^{-1}c^{-1}q^{1-k}e^{i\phi} \end{matrix};\ q\right]$$

$$+ \frac{(t)_\infty (at/c)_\infty (b^2c^2)_\infty \left|(ae^{i\phi})_\infty (be^{i\theta})_\infty (cte^{i\theta})_\infty (bca^{-1}te^{i\phi})_\infty\right|^2}{(ab)_\infty (ac)_\infty (bc)_\infty (a/c)_\infty (b^2c/a)_\infty (bc^2t/a)_\infty (bc/t)_\infty \left|(te^{i\theta+i\phi})_\infty (te^{i\theta-i\phi})_\infty\right|^2}$$

$$\cdot \sum_{k=0}^{\infty} \frac{(-t)_k (t\sqrt{q})_k (-t\sqrt{q})_k (bc^2t/a)_k \left|(te^{i\theta+i\phi})_k (te^{i\theta-i\phi})_k\right|^2}{(q)_k (qt/bc)_k (qt^2)_k (at/c)_k \left|(cte^{i\theta})_k (bca^{-1}te^{i\phi})_k\right|^2} q^k$$

$$\cdot {}_{10}\phi_9\left[\begin{matrix} bc^2a^{-1}tq^{k-1}, q\sqrt{\cdot}, -q\sqrt{\cdot}, tq^k, bctq^{k-1}, ca^{-1}tq^k, bca^{-1}e^{i\theta}, \\ \sqrt{\cdot}, -\sqrt{\cdot}, bc^2/a, cq/a, bc, ctq^ke^{-i\theta}, \end{matrix} \right.$$

$$\left. \begin{matrix} bca^{-1}e^{-i\theta}, ce^{i\phi}, ce^{-i\phi} \\ ctq^ke^{i\theta}, bca^{-1}tq^ke^{-i\phi}, bca^{-1}tq^ke^{i\phi} \end{matrix} ; q \right]$$

$$+ \frac{(t)_\infty (ct/a)_\infty (b^2c^2)_\infty \left|(ce^{i\phi})_\infty (bca^{-1}e^{i\theta})_\infty (ate^{i\theta})_\infty (bte^{i\phi})_\infty\right|^2}{(ac)_\infty (bc)_\infty (bc^2/a)_\infty (b^2c/a)_\infty (c/a)_\infty (abt)_\infty (bc/t)_\infty \left|(te^{i\theta+i\phi})_\infty (te^{i\theta-i\phi})_\infty\right|^2}$$

$$\cdot \sum_{k=0}^{\infty} \frac{(-t)_k (t\sqrt{q})_k (-t\sqrt{q})_k (abt)_k \left|(te^{i\theta+i\phi})_k (te^{i\theta-i\phi})_k\right|^2}{(q)_k (qt/bc)_k (qt^2)_k (ct/a)_k \left|(ate^{i\theta})_k (bte^{i\phi})_k\right|^2} q^k$$

$$\cdot {}_{10}\phi_9\left[\begin{matrix} abtq^{k-1}, q\sqrt{\cdot}, -q\sqrt{\cdot}, tq^k, bctq^{k-1}, ac^{-1}tq^k, be^{i\theta}, be^{-i\theta}, ae^{i\phi}, ae^{-i\phi} \\ \sqrt{\cdot}, -\sqrt{\cdot}, ab, aq/c, bc, atq^ke^{-i\theta}, atq^ke^{i\theta}, btq^ke^{-i\phi}, btq^ke^{i\phi} \end{matrix} ; q \right].$$

By inspection it is obvious from (6.13) that if $0 \leqq t < 1$, $0 < q < 1$, and (4.9) holds with $\alpha, \beta > -1$, then $K_t(x,y;\ a,b,c,bc/a;\ q)$ and hence the Poisson kernel $P_t(x,y;\ a,b,c,bc/a;\ q)$ for the continuous q-Jacobi polynomials are positive when $-1 \leqq x,y \leqq 1$.

REFERENCES

[1] W. A. Al-Salam and A. Verma, *Some remarks on q-beta integrals*, Proc. Amer. Math. Soc., 85 (1982), pp. 360–362.

[2] G. E. Andrews, *q-identities of Auluck, Carlitz and Rogers*, Duke Math. J., 33 (1966), pp. 575–581.

[3] _____, *On q-analogues of the Watson and Whipple summations*, this Journal, 7 (1976), pp. 332–336.

[4] R. Askey and J. Wilson, *Some basic hypergeometric orthogonal polynomials that generalize Jacobi polynomials*, Mem. Amer. Math. Soc., 319 (1985).

[5] W. N. Bailey, *Generalized Hypergeometric Series*, Stechert-Hafner Service Agency, New York and London, 1964.

[6] _____, *A transformation of nearly-poised basic hypergeometric series*, J. Lond. Math. Soc., 22 (1947), pp. 237–240.

[7] _____, *Series of hypergeometric type which are infinite in both directions*, Quart. J. Math. (Oxford), 7 (1936), pp. 105–115.

[8] L. Carlitz, *Some formulas of F. H. Jackson*, Monatsch. für Math., 73 (1969), pp. 193–198.

[9] George Gasper and Mizan Rahman, *Positivity of the Poisson kernel for the continuous q-ultraspherical polynomials*, this Journal, 14 (1983), pp. 409–420.

[10] _____, *Product formulas of Watson, Bailey and Bateman types and positivity of the Poisson kernel for q-Racah polynomials*, this Journal, 15 (1984), pp. 768–789.

[11] George Gasper, *A convolution structure and positivity of a generalized translation operator for the continuous q-Jacobi polynomials*, Conference on Harmonic Analysis in Honor of Antoni Zygmund, Wadsworth International Group, Belmont, CA, pp. 44–59.

[12] M. E.-H. Ismail and J. Wilson, *Asymptotic and generating relations for the q-Jacobi and ${}_4\phi_3$ polynomials*, J. Approx. Theory, 36 (1983), pp. 43–54.

[13] Mizan Rahman, *The linearization of the product of continuous q-Jacobi polynomials*, Canad. J. Math., 33 (1981), pp. 961–987.

[14] D. B. Sears, *Transformations of basic hypergeometric functions of special type*, Proc. Lond. Math. Soc., (2) 52 (1951), pp. 467–483.

[15] _____, *On the transformation theory of basic hypergeometric functions*, Proc. Lond. Math. Soc., 53 (1951), pp. 158–180.

[16] L. J. Slater, *Generalized Hypergeometric Functions*, Cambridge Univ. Press, Cambridge, 1966.

[17] Arun Verma, *A quadratic transformation of a basic hypergeometric series*, this Journal, 11 (1980), pp. 425–427.

SIAM J. MATH. ANAL.
Vol. 17, No. 4, July 1986

016

INEQUALITIES AND NUMERICAL BOUNDS FOR ZEROS OF ULTRASPHERICAL POLYNOMIALS*

SHAFIQUE AHMED†, MARTIN E. MULDOON‡ AND RENATO SPIGLER§

Abstract. We improve previous results concerning the monotonicity in λ of $f(\lambda)x_{nk}^{(\lambda)}$ where $x_{nk}^{(\lambda)}$ is a positive zero of the ultraspherical polynomial $P_n^{(\lambda)}(x)$, and $f(\lambda)$ is a suitably chosen positive increasing function. The range of validity is extended to $-\frac{1}{2} \leqq \lambda \leqq \frac{3}{2}$ rather than $0 \leqq \lambda \leqq 1$. In a certain sense the results are the best obtainable by the methods used. Some new elementary bounds for the zeros are obtained and compared with known results. An inequality for $\partial(\log x_{nk}^{(\lambda)})/\partial\lambda$ is also derived.

Key words. ultraspherical polynomials, zeros, inequalities, monotonicity, Sturm comparison theorem

AMS(MOS) subject classifications. Primary 33A65; secondary 34C10

1. Introduction. The study of zeros of special functions is of interest not only because of its mathematical aspects but also due to its many applications. The zeros of classical orthogonal polynomials may be interpreted as equilibrium configurations of certain one-dimensional systems [3], [12]. For instance the x-zeros $x_{nk}^{(\lambda)}$, $k=1,2,\cdots,n$, of the ultraspherical polynomial $P_n^{(\lambda)}(x)$ can be thought of as the positions of equilibrium of $n(\geqq 2)$ unit electrical charges in the interval $(-1,1)$ in the field generated by two identical charges of magnitude $\lambda/2+1/4$ placed at 1 and -1 [12, pp. 140–142]. Thus, the study of the variation of $x_{nk}^{(\lambda)}$ with λ and of bounds for $x_{nk}^{(\lambda)}$ corresponds to the physical problem of how the position of equilibrium of the n unit charges varies with λ and how much they can be displaced by changing λ. The physical interpretation makes it clear that the positive zeros move to the left when λ is increased; this agrees with a theorem of Stieltjes [9], [12, p. 121].

In this paper we use Sturmian methods to improve some previously known results [2], [7], [8] concerning the monotonicity of $f(\lambda)x_{nk}^{(\lambda)}$. From this we are able to derive some new elementary bounds for the zeros. In many cases these bounds are sharper than existing elementary bounds and, in some cases, generally for the smallest positive zero, they are sharper than existing bounds involving zeros of Bessel functions.

2. A preliminary theorem. Although the Sturm comparison theorem [10] is almost 150 years old, its value in providing information about zeros of special functions has not always been appreciated fully; see, for example, the remarks in Watson's treatise [13, p. 517]. It is largely to G. Szegö that we owe the obervation that Sturm methods can provide quite sharp results; see [12, Chap. 6] and R. Askey's notes following [11] in Szegö's collected papers.

An old result, due to Stieltjes [9], [12, p. 121] asserts that each positive zero $x_{nk}^{(\lambda)}$ of $P_n^{(\lambda)}(x)$ *decreases* as λ increases, for n, k fixed. On the other hand, R. Spigler [8] proved essentially the following theorem.

*Received by the editors September 20, 1984. This work was supported partially by the Natural Sciences and Engineering Research Council of Canada and by Consiglio Nazionale delle Ricerche, Italy.

†Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, New York 12181-3590.

‡Department of Mathematics, York University, Downsview, Ontario M3J 1P3, Canada.

§Courant Institute of Mathematical Sciences, New York University, New York, New York 10012, on leave from the University of Padua, Italy.

THEOREM 2.1. *The function* $f(\lambda)x_{nk}^{(\lambda)}$ *increases with* $\lambda, \lambda \in I$, *where* I *is some interval, for every function* $f(\lambda)$ *which satisfies* $f(\lambda)>0$, $f'(\lambda)>0$, $f'(\lambda)$ *continuous for* $\lambda \in I$ *and*

$$(2.1)\qquad (f^2-x^2)\left[2(n+\lambda)^2 f'-(2n+1)f\right]f+2x^2(n+\lambda)(f^2-x^2)$$
$$+ff'(1+2\lambda-2\lambda^2)(f^2+x^2)+x^2ff'\geqq 0, \qquad 0<x<f(\lambda).$$

In [8] only the case $I=(0,1)$ was considered. However, an examination of the proof reveals that we may take I to be any interval included in $[-\frac{1}{2}, \frac{3}{2}]$ possibly including one or both of its endpoints.

Theorem 2.1 was proved in [8] by applying a version of the Sturm comparison theorem [4] to the differential equation obtained by applying the scaling $x \to x/f(\lambda)$ to the equation satisfied by

$$u_\lambda(x)=(1-x^2)^{\lambda/2+1/4}P_n^{(\lambda)}(x).$$

The coefficient $\psi_\lambda(x)$ in the scaled equation

$$(2.2)\qquad u_\lambda''+\psi_\lambda(x)u_\lambda=0$$

was shown to be a monotonically decreasing function of λ using the condition (2.1). Spigler [8] called a function $f(\lambda)$ "acceptable" if it satisfies the hypotheses placed on f in Theorem 2.1 (in the special case where $I=(0,1)$). He then used (2.1) to derive the sufficient condition

$$(2.3)\qquad \frac{f'(\lambda)}{f(\lambda)} \geqq \frac{2n+1}{2(n+\lambda)^2}, \qquad 0<\lambda<1,$$

for the increase of $f(\lambda)x_{nk}^{(\lambda)}$ on $0<\lambda<1$ for an acceptable f. The results [8]

$$(2.4)\qquad \left|\frac{\partial\left(\log x_{nk}^{(\lambda)}\right)}{\partial\lambda}\right| \leqq \frac{f'(\lambda)}{f(\lambda)}, \qquad 0<\lambda<1,$$

and

$$(2.5)\qquad 1<\frac{x_{nk}^{(\lambda)}}{x_{nk}^{(\lambda+\varepsilon)}}<\frac{f(\lambda+\varepsilon)}{f(\lambda)}, \qquad 0<\lambda<\lambda+\varepsilon<1,$$

valid *for every acceptable* f showed that the most desirable f is one which minimizes the ratio $f'(\lambda)/f(\lambda)$. Correspondingly, the best results of the type (2.4) and (2.5) in [8] were obtained by taking the equality sign in (2.3). This leads, apart from a multiplicative constant, to

$$(2.6)\qquad f(\lambda)=\exp\left\{\frac{(2n+1)\lambda}{2n(n+\lambda)}\right\}.$$

The inequality

$$(2.7)\qquad \frac{x_{nk}^{(\lambda)}}{x_{nk}^{(\lambda+\varepsilon)}}<1+\frac{\varepsilon}{\lambda}$$

obtained in [7] corresponds to (2.5) with $f(\lambda)=\lambda$. However the right-hand side here becomes infinite as $\lambda \to 0$. Following a suggestion of R. Askey, S. Ahmed [2] considered functions of the form $(\lambda+m)^{1/2}$. Our results generalize those of [2].

In [8], use was made of the *sufficient* condition (2.3) for a function satisfying the other hypotheses in Theorem 2.1 to satisfy (2.1) also. Here we use instead (3.7) below which is, in a certain sense, both necessary and sufficient. In this way, we obtain inequalities of the type (2.4) and (2.5) which are the "best" that can be obtained by "the Sturm method plus scaling" in the sense that we discover the "acceptable" f's for which $f'(\lambda)/f(\lambda)$ is smallest on $-\frac{1}{2} \leqq \lambda \leqq \frac{3}{2}$. (Here we use the word "acceptable" in a wider sense than Spigler [8] who considered only $0 < \lambda < 1$.)

3. The best acceptable function. Our main goal here is to prove the following result.

THEOREM 3.1. *For $n \geqq 2$ let $x_{nk}^{(\lambda)}$ denote the kth x-zero in decreasing order of $P_n^{(\lambda)}(x)$. Then, for $k = 1, \cdots, [n/2]$,*

$$[2n^2+1+2\lambda(2n+1)]^{1/2} x_{nk}^{(\lambda)}$$

increases as λ increases, $-\frac{1}{2} \leqq \lambda \leqq \frac{3}{2}$.

Proof. It is simply a matter of checking that the condition (2.1), with equality sign, is satisfied for $0 \leqq x \leqq f(\lambda)$, $-\frac{1}{2} \leqq \lambda \leqq \frac{3}{2}$, where

$$f(\lambda) = [2n^2+1+2\lambda(2n+1)]^{1/2}. \tag{3.1}$$

To see this, we divide (2.1) by $[f(\lambda)]^4$ and obtain, after some algebraic manipulation,

$$\frac{f'(\lambda)}{f(\lambda)} \geqq F\left(\frac{x^2}{f^2}\right), \qquad 0 < x < f(\lambda), \tag{3.2}$$

where

$$F(u) = \frac{(1-u)[2(n+1)-2u(n+\lambda)]}{2(n+\lambda)^2(1-u)+(1+2\lambda-2\lambda^2)(1+u)+u}. \tag{3.3}$$

It is clear that the denominator here is positive for $0 < u < 1$, $-\frac{1}{2} \leqq \lambda \leqq \frac{3}{2}$.

In order that (3.2) hold *for every* x in $(0, f(\lambda))$ it is necessary and sufficient that

$$f'(\lambda)/f(\lambda) \geqq \sup_{0<u<1} F(u). \tag{3.4}$$

However we can show that

$$\sup_{0<u<1} F(u) = F(0). \tag{3.5}$$

To prove (3.5) we note that it is equivalent to

$$\frac{a-(a+b)u+bu^2}{A-Bu} \leqq \frac{a}{A}, \qquad 0 < u < 1 \tag{3.6}$$

where $a = 2n+1$, $b = 2(n+\lambda)$,

$$A = 2(n+\lambda)^2+1+2\lambda-2\lambda^2, \qquad B = -2\left[1+\lambda-\lambda^2-(n+\lambda)^2\right]$$

and it is easy to see that a necessary and sufficient condition for this is that $A \geqq B$ or that $-\frac{1}{2} \leqq \lambda \leqq \frac{3}{2}$. Thus for these values of λ, (3.4) is equivalent to

$$\frac{f'(\lambda)}{f(\lambda)} \geqq \frac{2n+1}{2(n+\lambda)^2+1+2\lambda-2\lambda^2} \tag{3.7}$$

and it is easy to see that this holds with the equality sign when $f(\lambda)$ is given by (3.1). This completes the proof of Theorem 3.1.

The method of proof shows that the choice (3.1) for $f(\lambda)$ is, apart from constant factors, the one which *minimizes* $f'(\lambda)/f(\lambda)$ subject to the hypotheses of Theorem 2.1 applied to the interval $I=[-\frac{1}{2},\frac{3}{2}]$ and so it is the best "acceptable" function for this interval.

From Theorem 3.1 we get

COROLLARY 3.1. *If* $n\geqq 2$, $-\frac{1}{2}\leqq\lambda<\lambda+\varepsilon<\frac{3}{2}$, $k=1,2,\ldots[n/2]$, *we have*

$$1<\frac{x_{nk}^{(\lambda)}}{x_{nk}^{(\lambda+\varepsilon)}}<\left[1+\frac{\varepsilon}{\lambda+(2n^2+1)/[2(2n+1)]}\right]^{1/2}.$$

We also have, as in [8], from (2.4) and (3.7):

COROLLARY 3.2. *A "best" estimate for the derivative of* $x_{nk}^{(\lambda)}$ *is given by*

$$\left|\frac{\partial\left(\log x_{nk}^{(\lambda)}\right)}{\partial\lambda}\right|<\frac{2n+1}{2(2n+1)\lambda+2n^2+1}$$

where $-\frac{1}{2}\leqq\lambda\leqq\frac{3}{2}$, $k=1,2,\cdots,[n/2]$.

4. Numerical bounds for $x_{nk}^{(\lambda)}$, $0<\lambda<1$. The decreasing character of $x_{nk}^{(\lambda)}$, as a function of λ, enables us to provide *upper and lower bounds* for this quantity whenever the zeros of two other ultraspherical polynomials $P_n^{(\lambda_1)}(x)$ and $P_n^{(\lambda_2)}(x)$, with $\lambda_1<\lambda<\lambda_2$, are known. When $\lambda_1=0$, $\lambda_2=1$ we have the Chebyshev polynomials; in this way Stieltjes was able to show that

$$\cos\frac{k\pi}{n+1}<x_{nk}^{(\lambda)}<\cos\frac{(2k-1)\pi}{2n}, \tag{4.1}$$

for $0<\lambda<1$ [9], [12, p. 122].

The monotonic increasing character of $f(\lambda)x_{nk}^{(\lambda)}$, where $f(\lambda)$ is given by (3.1), enables us to get sharper results. In fact we have

$$\frac{f(\lambda_1)}{f(\lambda)}x_{nk}^{(\lambda_1)}<x_{nk}^{(\lambda)}<\frac{f(\lambda_2)}{f(\lambda)}x_{nk}^{(\lambda_2)} \tag{4.2}$$

where $-\frac{1}{2}\leqq\lambda_1<\lambda<\lambda_2\leqq\frac{3}{2}$, $k=1,2,\cdots,[n/2]$. In particular with $\lambda_1=0$, $\lambda_2=1$ we get the following result.

THEOREM 4.1. *For* $0<\lambda<1$ *and* $k=1,2,\cdots,[n/2]$ *we have*

$$x_{nk}^{(\lambda)}<\left[\frac{2n^2+4n+3}{2n^2+1+2\lambda(2n+1)}\right]^{1/2}\cos\frac{k\pi}{n+1} \tag{4.3}$$

and

$$x_{nk}^{(\lambda)}>\left[\frac{2n^2+1}{2n^2+1+2\lambda(2n+1)}\right]^{1/2}\cos\frac{(2k-1)\pi}{2n}. \tag{4.4}$$

G. Szegö [12, §6.6] summarizes the various known elementary bounds for $x_{nk}^{(\lambda)}$, including some obtained by Sturm methods. The best such *upper* bound, among those available in [12] is

$$x_{nk}^{(\lambda)}<\cos\left[\frac{k-(1-\lambda)/2}{n+\lambda}\pi\right],\qquad 0<\lambda<1, \tag{4.5}$$

and the best such *lower* bounds are

$$x_{nk}^{(\lambda)} > \cos\frac{k\pi}{n+1}, \qquad \frac{1}{2} \leqq \lambda < 1, \tag{4.6}$$

and

$$x_{nk}^{(\lambda)} > \cos\left\{\frac{(k+\lambda-1/2)\pi}{n+2\lambda}\right\}, \qquad 0 < \lambda \leqq \frac{1}{2}. \tag{4.7}$$

In the special case $\lambda = \frac{1}{2}$ (Legendre polynomials) some calculations which we performed in the range $n = 2$ to 15 indicate that our lower bound (4.4) is quite sharp and generally better than (4.6) and (4.7) for all except the first (largest) one or two zeros, while our upper bound (4.3) is better than (4.5) for at most the last (smallest) zero. Thus, for example, our bounds lead to

$$0.20111 < x_{15,7}^{(1/2)} < 0.20126 \tag{4.8}$$

while (4.5) and (4.6) lead to

$$0.19509 < x_{15,7}^{(1/2)} < 0.20130.$$

We note that (4.1) would have given only

$$0.195090 < x_{15,7}^{(1/2)} < 0.207912.$$

Calculations for other values of λ $(0 < \lambda < 1)$ confirm the pattern; our lower bounds are better than (4.6) and (4.7) for most zeros except the largest while our upper bounds are better than (4.5) only for the smallest zeros. There does not appear to be any easy way to decide a priori on the relative sharpness of these elementary bounds. In Table 1 we provide some further numerical comparison of our bounds with (4.5) to (4.7).

TABLE 1

n	k	L	L_s	U	U_s
2	1	0.56695*	0.50000	0.58248*	0.58779
3	1	0.74032*	0.70711	0.79663	0.78183
4	1	0.81893*	0.80902	0.89149	0.86603
4	2	0.33921*	0.30902	0.34052*	0.34202
5	1	0.86257	0.86603	0.93972	0.90963
5	2	0.53310*	0.50000	0.54254	0.54064
6	1	0.88993	0.90100	0.96667	0.93502
6	2	0.65147*	0.62349	0.66896	0.66312
6	3	0.23846*	0.22252	0.23875*	0.23932

The columns headed L and U give the (lower and upper) bounds for $x_{nk}^{(1/2)}$ *obtained using our formulas* (4.3) *and* (4.4). *The columns headed* L_s *and* U_s *give the bounds obtained by using* (4.5) *and* (4.6) (*or* (4.7)). *An asterisk indicates the cases where our bounds are better.*

We can hardly expect our bounds to compare as well with those involving nonelementary functions. Szegö [11] showed that for $0 < \lambda < 1$, $k = 1, 2, \cdots, [n/2]$,

$$x_{nk}^{(\lambda)} < \cos\frac{j_{\lambda-1/2,k}}{\left[(n+\lambda)^2 + (1-4/\pi^2)\lambda(1-\lambda)\right]^{1/2}} \tag{4.9}$$

where $j_{\lambda-1/2,k}$ is the kth positive zero of the Bessel function $J_{\lambda-1/2}(x)$, while L. Gatteschi [5] provided the lower bound

$$x_{nk}^{(\lambda)} > \cos \frac{j_{\lambda-1/2,k}}{\left[(n+\lambda)^2+\lambda(1-\lambda)/3\right]^{1/2}} \tag{4.10}$$

for the same range. Some calculations indicate that our elementary bounds (4.3), (4.4) are sometimes sharper than (4.9), (4.10) but only in the case of the smallest zeros. Thus, for example, (4.3) and (4.4) give

$$0.14885 < x_{10,5}^{(1/2)} < 0.14889$$

while (4.9) and (4.10) give

$$0.148788 < x_{10,5}^{(1/2)} < 0.149204.$$

5. Bounds outside the range $0<\lambda<1$. It is clear from (4.2) that we also have, for $1<\lambda\leqq\frac{3}{2}$,

$$x_{nk}^{(\lambda)} > \left[\frac{2n^2+4n+3}{2n^2+1+2\lambda(2n+1)}\right]^{1/2} \cos\frac{k\pi}{n+1} \tag{5.1}$$

and for $1\leqq\lambda<\frac{3}{2}$

$$x_{nk}^{(\lambda)} < \left[\frac{2n^2+6n+4}{2n^2+1+2\lambda(2n+1)}\right] x_{nk}^{(3/2)}. \tag{5.2}$$

The corresponding inequalities for $-\frac{1}{2}\leqq\lambda\leqq 0$ are

$$x_{nk}^{(\lambda)} > \left[\frac{2n^2-2n}{2n^2+1+2\lambda(2n+1)}\right]^{1/2} x_{nk}^{(-1/2)}, \qquad -\frac{1}{2}<\lambda\leqq 0 \tag{5.3}$$

and

$$x_{nk}^{(\lambda)} < \left[\frac{2n^2+1}{2n^2+1+2\lambda(2n+1)}\right]^{1/2} \cos\frac{(2k-1)\pi}{2n}, \qquad -\frac{1}{2}\leqq\lambda<0. \tag{5.4}$$

All of the above inequalities hold for $k=1,2,\cdots,[n/2]$. It is noteworthy that (5.1) and (5.4) are reversed forms of (4.3) and (4.4). Laforgia [6] has shown that inequalities (4.9) and (4.10), which become equalities for $\lambda=1$, are reversed for $\lambda>1$ so it is of interest to compare the elementary lower bound (5.1) with the reversed form of (4.10) in the range $1<\lambda\leqq\frac{3}{2}$. Some calculations which we have performed in the case $n=10$ and various values of λ indicate that the elementary bound (5.1) is sharper only in the case of the smallest zero ($k=5$ in this example). The upper bound (5.2) is not elementary but it involves the zeros of $P_n^{(3/2)}(x)$ which are the same as those of the *derivative* $P'_{n+1}(x)$ *of the Legendre polynomial* and some of these have been tabulated [1, p. 920].

In particular, the inequality (5.3) and, more especially the elementary inequality (5.4), are noteworthy in that they provide simple bounds corresponding to *negative values of* λ. There do not appear to be many such bounds in the literature.

6. Concluding remarks. 1. Since Laforgia's reversed forms of (4.9), (4.10) are valid for *all* $\lambda>1$, the question arises as to whether some of our monotonicity properties or

elementary bounds can be extended to all values of $\lambda > 1$. We have seen that our method of proof *cannot* furnish such an extension since inequality (3.6) fails to hold for $\lambda > 3/2$. However the example

$$x_{21}^{(\lambda)} = [2(\lambda+1)]^{-1/2}$$

shows that, for $\varepsilon > 0$, $(\lambda+1-\varepsilon)^{1/2} x_{21}^{(\lambda)}$ increases on the interval $-1+\varepsilon < \lambda < \infty$. Examining the proof in §3 we see that if we take $\lambda > \frac{3}{2}$ the inequality (2.1) breaks down for x close to $f(\lambda)$. The condition $0 < x < f(\lambda)$ arises from the fact that the positive zeros lie in $(0,1)$. If we have a better upper bound for a zero say $0 < x_{nk}^{(\lambda)} < x_0 < 1$ then we would need the condition (2.1) only for $0 < x < x_0 f(\lambda)$ with a corresponding enlargement of the set of values of λ for which our results hold. Since we have many such upper bounds including those proved here, our results can in fact be improved in this way though it does not seem to be possible to get them for *all* positive λ.

2. In the case of $x_{21}^{(\lambda)}$, not only do our methods not give the whole λ-interval, they do not in fact give the factor $(\lambda+1-\varepsilon)^{1/2}$ for small $\varepsilon > 0$. Using our method we need (2.1) to hold as $x \to 0^+$ i.e., we need

$$[(2n+1)(n+2\lambda)-(n-1)] f'(\lambda) \geqq (2n+1) f(\lambda). \tag{6.1}$$

If we let $f(\lambda) = (\lambda+\beta)^{1/2}$, we see that (6.1) requires us to have $\beta \leqq \frac{9}{10}$.

3. In [8] it was shown that $\lambda^{\alpha} x_{nk}^{(\lambda)}$ increases with λ, $0 < \lambda < 1$, provided

$$\alpha \geqq \max_{n \geqq 1} \frac{2n+1}{8n} = \frac{3}{8}.$$

Actually, since only the cases $n \geqq 2$ arise, it could be deduced from the work in [8] that this holds for

$$\alpha \geqq \max_{n \geqq 2} \frac{2n+1}{8n} = \frac{5}{16}.$$

However the results of the present paper, in particular the condition (3.7), show that it is sufficient to have

$$\alpha \geqq \max \frac{2n+1}{2(2n+1)+2n^2+1} = \frac{5}{19}.$$

4. $P_4^{(\lambda)}(x)$ is a constant multiple of

$$4(\lambda+2)(\lambda+3)x^4 - 12(\lambda+2)x^2 + 3.$$

Hence $[x_{41}^{(\lambda)}]^2$ and $[x_{42}^{(\lambda)}]^2$ are given by

$$\left[3 \pm \sqrt{9 - 3(\lambda+3)/(\lambda+2)}\,\right] / [2(\lambda+3)].$$

From this it follows that $(\lambda+3)^{1/2} x_{41}^{(\lambda)}$ *increases* while $(\lambda+3)^{1/2} x_{42}^{(\lambda)}$ *decreases* on $(-3/2, \infty)$. Thus, it would be of interest to obtain monotonicity results like Theorem 3.1 in which the square root factor depends on k as well as on n. Such results would lead to elementary bounds of the form (4.3), (4.4) in which the square root factors as well as the arguments of the cosine terms depend on k.

REFERENCES

[1] M. ABRAMOWITZ AND I. A. STEGUN, eds., *Handbook of Mathematical Functions*, Applied Mathematics Series, 55, National Bureau of Standards, Washington, DC, 1964.

[2] S. AHMED, *On the zeros of orthogonal polynomials*, Abstracts Amer. Math. Soc., 3 (1982), p. 339.

[3] S. AMHED, M. BRUSCHI, F. CALOGERO, M. A. OLSHANETSKY AND A. M. PERELOMOV, *Properties of the zeros of the classical polynomials and of the Bessel functions*, Nuovo Cimento, 49B (1979), pp. 173–199.

[4] S. AMHED, A. LAFORGIA AND M. E. MULDOON, *On the spacing of the zeros of some classical orthogonal polynomials*, J. London Math. Soc. (2), 25 (1982), pp. 246–252.

[5] L. GATTESCHI, *Una nuova disuguaglianza per gli zeri dei polinomi di Jacobi*, Atti Accad. Sci. Torino, Cl. Sci. Fis. Mat. Natur., 103 (1968–69), pp. 259–265.

[6] A. LAFORGIA, *Sugli zeri delle funzioni di Bessel*, Calcolo, 17 (1980), pp. 211–220.

[7] _____, *A monotonic property for the zeros of ultraspherical polynomials*, Proc. Amer. Math. Soc., 83 (1981), pp. 757–758.

[8] R. SPIGLER, *On the monotonic variation of the zeros of ultraspherical polynomials with the parameter*, Canad. Math. Bull., 27 (1984), pp. 472–477.

[9] T. J. STIELTJES, *Sur les racines de l'equation* $X_n = 0$, Acta Math., 9 (1886), pp. 385–400; Oeuvres Complètes, vol. 2, pp. 73–88.

[10] CH. STURM, *Mémoire sur les équations différentielles du second ordre*, J. Math. Pures Appl., 1 (1836), pp. 106–186.

[11] G. SZEGÖ, *Inequalities for the zeros of Legendre polynomials and related functions*, Trans. Amer. Math. Soc., 39 (1936), pp. 1–17; Collected papers, vol. 2, pp. 593–609.

[12] _____, *Orthogonal Polynomials*, 4th ed., AMS Colloquium Publications 23, American Mathematical Society, Providence, RI, 1975.

[13] G. N. WATSON, *A Treatise on the Theory of Bessel Functions*, 2nd ed., Cambridge Univ. Press, Cambridge, 1944.

SIAM J. MATH. ANAL.
Vol. 17, No. 4, July 1986

017

A NOTE TO THE PAPER OF AHMED, MULDOON AND SPIGLER*

ÁRPÁD ELBERT† AND ANDREA LAFORGIA‡

Abstract. We give a simplified proof of a theorem proved in [*Inequalities and numerical bounds for zeros of ultraspherical polynomials*, this Journal, 17 (1986) pp. 1000–1007].

Key words. zeros of ultraspherical polynomials, Sturm comparison theorem

AMS(MOS) subject classifications. Primary 33A45; secondary 34C10

S. Ahmed, M. E. Muldoon and R. Spigler proved the following result [2].

THEOREM. *For $n \geqq 2$ let $x_{nk}^{(\lambda)}$ denote the kth x-zero in decreasing order of $P_n^{(\lambda)}(x)$. Then for $k = 1, 2, \cdots, [n/2]$*

$$[2n^2+1+2\lambda(2n+1)]^{1/2} x_{nk}^{(\lambda)}$$

increases as λ increases, $-\frac{1}{2} \leqq \lambda \leqq \frac{3}{2}$.

The proof of the theorem given in [2] was based on a result due to Spigler, proved as a consequence of a version of the Sturm comparison theorem given in [1].

The aim of this note is to provide a shorter and easier proof of the theorem above.

Proof. The function $u(x) = (1-x^2)^{\lambda/2+1/4} P_n^{(\lambda)}(x)$ satisfies the differential equation [3, p. 81]

$$\frac{d^2u}{dx^2} + p(\lambda, x)u = 0,$$

where

$$p(\lambda, x) = \frac{(n+\lambda)^2}{1-x^2} + \frac{1/2+\lambda-\lambda^2+x^2/4}{(1-x^2)^2}.$$

Let $t = f(\lambda)x$, where

$$f(\lambda) = [2n^2+1+2\lambda(2n+1)]^{1/2}.$$

Then the function $U(t) = u(x)$ satisfies the differential equation

$$\frac{d^2U}{dt^2} + \tilde{P}(\lambda, t)U = 0,$$

with

$$\tilde{P}(\lambda, t) = [f(\lambda)]^{-2} p\left(\lambda, \frac{t}{f(\lambda)}\right).$$

*Received by the editors April 5, 1985. This work was supported by the Consiglio Nazionale delle Ricerche of Italy.

†Mathematical Institute of the Hungarian Academy of Sciences, Budapest, P.f. 428, 1376 Hungary.

‡Department of Mathematics, University of Torino, Via Carlo Alberto, 10 10123 Torino, Italy.

The authors in [2] have this result. Making the substitution

$$t^2=\sigma, \qquad f^2=\varphi=2n^2+1+2\lambda(2n+1)$$

we get

$$P(\lambda,\varphi,\sigma)=\tilde{P}(\lambda,t)=\frac{(n+\lambda)^2}{\varphi-\sigma}+\frac{\varphi(1/2+\lambda-\lambda^2)+\sigma/4}{(\varphi-\sigma)^2}.$$

As in [2] we intend to apply the version of the Sturm comparison theorem proved in [1]. To this end we need to show that the function $P(\lambda,\varphi,\sigma)$ defined above decreases as λ increases, $-\frac{1}{2}\leqq\lambda\leqq\frac{3}{2}$. Since

$$\frac{dP}{d\lambda}=\frac{\partial P}{\partial\lambda}+\frac{\partial P}{\partial\varphi}\varphi', \qquad 0\leqq\sigma<\varphi$$

we get

$$(\varphi-\sigma)^3\frac{dP}{d\lambda}=A\sigma^2+B\sigma+C, \tag{1}$$

where

$$A=2(n+\lambda), \qquad B=-(4n+2\lambda+1)\varphi+\left[\lambda^2-\lambda-1+(n+\lambda)^2\right]\varphi',$$
$$C=(2n+1)\varphi^2+\left[\lambda^2-\lambda-1/2-(n+\lambda)^2\right]\varphi\varphi'.$$

The quadratic polynomial (1) is strictly convex.

Therefore, in order to show that $dP/d\lambda\leqq 0$ we have to check the sign of this polynomial only at $\sigma=0$ and $\sigma=\varphi$. By the definition of σ we obtain $C=0$. At $\sigma=\varphi$ the polynomial assumes the value

$$\varphi(2n+1)(4\lambda^2-4\lambda-3)=\varphi(2n+1)(2\lambda+1)(2\lambda-3)$$

which is clearly negative if $-\frac{1}{2}<\lambda<\frac{3}{2}$.

The proof of the theorem is complete.

Acknowledgment. We are indebted to Professor Richard A. Askey for the careful reading of the manuscript.

REFERENCES

[1] S. AHMED, A. LAFORGIA AND M. E. MULDOON, *On the spacing of the zeros of some classical orthogonal polynomials*, J. London Math. Soc. (2), 25 (1982), pp. 246–252.

[2] S. AHMED, M. E. MULDOON AND R. SPIGLER, *Inequalities and numerical bounds for zeros of ultraspherical polynomials*, this Journal, 17 (1986), pp. 1000–1007.

[3] G. SZEGŐ, *Orthogonal Polynomials*, 4th ed., AMS Colloquium Publications 23, American Mathematical Society, Providence, RI, 1975.

SIAM J. MATH. ANAL.
Vol. 17, No. 4, July 1986

018

A NEW PROOF OF WATSON'S PRODUCT FORMULA FOR LAGUERRE POLYNOMIALS VIA A CAUCHY PROBLEM ASSOCIATED WITH A SINGULAR DIFFERENTIAL OPERATOR*

C. MARKETT†

Abstract. Watson's product formula for Laguerre polynomials L_n^α for $\alpha > -\frac{1}{2}$ as well as its limiting case for $\alpha = -\frac{1}{2}$ have served for defining a generalized translation operator [1]. In this paper we proceed just the other way. The Laguerre translation operator will be introduced as the solution of an associated Cauchy problem for each $\alpha \geqq -\frac{1}{2}$. A main task consists in finding the corresponding Riemann function explicitly. The Laguerre product formula then follows as a corollary. The paper concludes with a comparison between the translation operators associated with the Laguerre series and the Hankel transform.

Key words. Bessel function, Cauchy problem, generalized translation operator, Hankel transform, Laguerre polynomial, product formula, Riemann function, Sturm–Liouville differential equation

AMS(MOS) subject classifications. Primary 33A65, 35C10

1. Introduction, survey of approach. Let for $\alpha > -1$, $x \geqq 0$, and $n \in \mathbb{P} = \{0, 1, 2, \cdots\}$ the Laguerre polynomials, Laguerre functions, and the Bessel functions be defined by

$$L_n^\alpha(x) = \sum_{k=0}^{n} \binom{n+\alpha}{n-k} \frac{(-x)^k}{k!},$$

$$y_n^\alpha(x) = e^{-x^2/2} L_n^\alpha(x^2)/L_n^\alpha(0),$$

$$j_\alpha(x) = 2^\alpha \Gamma(\alpha+1) x^{-\alpha} J_\alpha(x) = \sum_{k=0}^{\infty} \frac{\Gamma(\alpha+1)(-1)^k (x/2)^{2k}}{k!\Gamma(k+\alpha+1)},$$

respectively. The product formula for Laguerre polynomials due to Watson [32] can be written in the form

$$y_n^\alpha(x) y_n^\alpha(t) = \frac{\Gamma(\alpha+1)}{\Gamma(\alpha+1/2)\Gamma(1/2)} \int_0^\pi y_n^\alpha\left([x^2+t^2+2xt\cos\theta]^{1/2}\right) \cdot j_{\alpha-1/2}(xt\sin\theta)\sin^{2\alpha}\theta\, d\theta \qquad (\alpha > -\tfrac{1}{2},\ x, t \geqq 0). \tag{1.1}$$

The limiting formula for α tending to $-\frac{1}{2}$,

$$y_n^{-1/2}(x) y_n^{-1/2}(t) = \frac{1}{2}\left\{ y_n^{-1/2}(|x-t|) + y_n^{-1/2}(x+t) \right\} - \frac{xt}{2} \int_0^\pi y_n^{-1/2}\left([x^2+t^2+2xt\cos\theta]^{1/2}\right) J_1(xt\sin\theta)\, d\theta, \tag{1.2}$$

was first proved by Boersma, cf. [23] for his proof and another proof by the author.

*Received by the editors March 16, 1983. This work was supported by the Deutsche Forschungsgemeinschaft under grant Go 261/5-1.

†Lehrstuhl A für Mathematik, Rheinisch-Westfälische Technische Hochschule Aachen, Federal Republic of Germany.

Watson proved the identity (1.1) by special function arguments using the generating functions of the Laguerre polynomials and their products as well as an integral representation of Bessel functions due to Gegenbauer. Later, a group theoretical proof for $\alpha \in \mathbb{P}$ was given by Peetre [24]: If one associates the Weyl transform with the k-dimensional Heisenberg group, the so-called twisted convolution of two functions is introduced in such a way that the convolution is mapped by the Weyl transform onto an ordinary product of the Weyl transforms of the functions. Peetre showed that the restriction of the Weyl transform to rotation invariant functions is the Laguerre transform with $\alpha = k-1$, and the restriction of the twisted convolution is the Laguerre convolution. This immediately yields the product formula.

Another proof is due to Koornwinder [20] who deduces the product formula for $\alpha > 0$ from his Laguerre addition formula by integration. The assertion for $-\frac{1}{2} < \alpha \leqq 0$ then follows by analytic continuation with respect to α. Koornwinder obtains his addition formula as a limit of the addition formula for the so-called disk polynomials which for its part was found in case $\alpha \in \mathbb{N}$ independently by Šapiro [28] and Koornwinder [16] by interpreting the disk polynomials as spherical functions on the homogeneous spaces $SU(\alpha+2)/SU(\alpha+1)$. Again, analytic continuation leads to the general formula. Thus one may say that Koornwinder's proof of (1.1) also has a group theoretical source.

Each of these proofs is of its own interest since each illustrates a different aspect of special function theory. The main purpose of the present paper is to give another analytic proof for the Laguerre product formula which works for each $\alpha \geqq -\frac{1}{2}$ at once. Our starting point is the fact that the Laguerre functions y_n^α are the eigenfunctions of the Sturm–Liouville differential equation (S.-L. DE)

$$\frac{d}{dx}\left[x^{2\alpha+1}\frac{d}{dx}u(x)\right]+(\lambda x^{2\alpha+1}-x^{2\alpha+3})u(x)=0 \qquad (0 \leqq x < \infty) \tag{1.3}$$

for the eigenvalues $\lambda_n = 4n+2\alpha+2$, satisfying the boundary conditions

$$u(0)=1, \qquad \int_0^\infty |u(x)|^2 x^{2\alpha+1}dx < \infty.$$

DE(1.3) follows from the more familiar S.-L. DE for the Laguerre polynomials [30]

$$\frac{d}{dx}\left[e^{-x}x^{\alpha+1}\frac{d}{dx}L_n^\alpha(x)\right]+ne^{-x}x^\alpha L_n^\alpha(x)=0 \qquad (n \in \mathbb{P})$$

by an appropriate transformation. In operational notation it can also be written as

$$D_{q,x}^\alpha(u,x)+\lambda u(x)=0, \qquad D_{q,x}^\alpha = \frac{d^2}{dx^2}+\frac{2\alpha+1}{x}\frac{d}{dx}-q(x) \tag{1.4}$$

with the potential function given by $q(x)=x^2$. The Laguerre translation $T_{L,t}^\alpha(f,x)$ of a sufficiently smooth function $f(x)$ can then be introduced as the solution $u(x,t)$ of the Cauchy problem

$$\begin{aligned} &\left(D_{q,x}^\alpha - D_{q,t}^\alpha\right)u(x,t)=0 \qquad (0<t\leqq x), \\ &u(x,0)=f(x), \qquad u_t(x,0)=\frac{\partial}{\partial t}u(x,t)\Big|_{t=0}=0 \qquad (x>0), \end{aligned} \tag{1.5}$$

where $q(x)=x^2$, and where $u(x,t)$ is extended symmetrically in x and t to $0<x<t$. By solving this problem we will establish the following kernel representation of the Laguerre translation (see also [1] and [12] with a slight change of notation).

THEOREM 1. *For $\alpha \geqq -\frac{1}{2}$ the solution $u(x,t)=T^{\alpha}_{L,t}(f;x)$ of the Cauchy problem* (1.5) *is given by*

$$T^{\alpha}_{L,t}(f;\,x)=\begin{cases}\displaystyle\int_0^{\infty} f(z)K^{\alpha}_L(x,t,z)z^{2\alpha+1}dz, & \alpha>-\frac{1}{2},\\[2ex] \displaystyle\frac{1}{2}\{f(|x-t|)+f(x+t)\}+\int_0^{\infty} f(z)K_L^{-1/2}(x,t,z)\,dz, & \alpha=-\frac{1}{2},\end{cases}$$

where the kernel functions K^{α}_L are

$$K^{\alpha}_L(x,t,z)=\begin{cases}\displaystyle\frac{\Gamma(\alpha+1)}{\Gamma(\alpha+1/2)\Gamma(1/2)}(xtz)^{-2\alpha}\rho^{2\alpha-1}j_{\alpha-1/2}(\rho), & \alpha>-\frac{1}{2},\\[2ex] \displaystyle-\frac{1}{4}xtz\,j_1(\rho), & \alpha=-\frac{1}{2}\end{cases}\tag{1.6}$$

for $|x-t|<z<x+t$, and $K^{\alpha}_L(x,t,z)=0$ elsewhere, with the notation

$$\begin{aligned}\rho(x,t,z)&=\frac{1}{2}\left(2[x^2t^2+x^2z^2+t^2z^2]-x^4-t^4-z^4\right)^{1/2}\\ &=\frac{1}{2}\left(\left[(x+t)^2-z^2\right]\left[z^2-(x-t)^2\right]\right)^{1/2}.\end{aligned}\tag{1.7}$$

The proof will be given in §5. For $\alpha\geqq 0$ the domain of the translation operator $T^{\alpha}_{L,t}$ can be extended to the weighted Lebesgue space

$$L^1_{w(2\alpha+1)}=\left\{f;\ \|f\|_{L^1_{w(2\alpha+1)}}=\int_0^{\infty}|f(x)|x^{2\alpha+1}dx<\infty\right\}$$

(cf. [12]). Applying Theorem 1 to the Laguerre functions y^{α}_n, $n\in\mathbb{P}$, one obtains

$$y^{\alpha}_n(x)y^{\alpha}_n(t)=T^{\alpha}_{L,t}(y^{\alpha}_n;\,x)\qquad\left(x,t>0,\alpha\geqq-\frac{1}{2}\right),$$

which provides the desired proof of the product formulas (1.1), (1.2).

The foregoing way of introducing generalized translation operators goes back to Delsarte [7], [8], who studied the Cauchy problem (1.5) in case $q(x)\equiv 0$. Many authors have continued and extended his work. For example, Levitan [22] and Povzner [25] considered the particular case $\alpha=-\frac{1}{2}$ of (1.5), imposing certain growth conditions on the potential q. Other generalizations were made by Leblanc, Hutson and Pym, as well as by Chébli; cf. [4] for references. For example, Chébli [5] proceeds from the partly more general differential operator

$$D^A_x=\frac{1}{A(x)}\frac{\partial}{\partial x}\left[A(x)\frac{\partial}{\partial x}\right]\qquad(0<x<\infty)$$

where, apart from a convexity property for $A(x)$, he assumes that $A\in C^2(0,\infty)$, $A(0)=0$, $A(x)>0$ for $x>0$, and $A'(x)/A(x)=\alpha/x+B(x)$, B being continuous at 0.

In order to solve a Cauchy problem of the form (1.5), Riemann's method can be used provided the corresponding Riemann function can be determined explicitly (cf. §2). Of course, the latter task is the main obstacle here. The interesting feature of this approach to a translation operator is that it can also cover variants of the classical translations which arise from slight variations of the potential function q in (1.5). In fact, once the Riemann function of a differential equation (1.5) is found for some fixed

potential q_0, it may be possible to handle the solutions of (1.5) for other q without knowing their Riemann functions explicitly, e.g., by estimating the Riemann functions in terms of the old one.

Braaksma and de Snoo [4] have proceeded this way, starting from the Cauchy problem (1.5) with $q_0 \equiv 0$ (cf. also [3]). Its solution defines the well-known "Hankel translation"

$$T^{\alpha}_{H,t}(f;\, x) = \begin{cases} \displaystyle\int_0^{\infty} f(z) K^{\alpha}_H(x,t,z) z^{2\alpha+1} dz, & \alpha > -\frac{1}{2}, \\ \displaystyle\frac{1}{2}\{f(|x-t|)+f(x+t)\}, & \alpha = -\frac{1}{2}, \end{cases} \tag{1.8}$$

where

$$K^{\alpha}_H(x,t,z) = \frac{\Gamma(\alpha+1)}{\Gamma(\alpha+1/2)\Gamma(1/2)} (xtz)^{-2\alpha} (\rho(x,t,z))^{2\alpha-1}$$

if $|x-t| < z < x+t$, and $K^{\alpha}_H(x,t,z) = 0$ elsewhere, with ρ given by (1.7). See [2], [6], [13], [15], also for applications in harmonic analysis of the Hankel transform. By varying the potential and the corresponding Riemann function (cf. (4.7), (4.9)) Braaksma and de Snoo derived estimates of the kernels of a class of generalized translations. This class does not contain the Laguerre translation whose potential is $q(x) = x^2$ since, among other conditions, the potentials q admitted in [4] have to vanish at infinity. So our final purpose is to determine the Riemann function associated with the Laguerre translation (cf. §4, in particular, (4.11), (4.13)) and thus to provide a new starting point ($q_0(x) = x^2$ instead of $q_0(x) \equiv 0$) for the same procedure. In a forthcoming paper it will be shown that in this way one can indeed obtain norm estimates for a class of translations in some neighborhood of the Laguerre case.

At this point let us make some further remarks concerning product formulas for orthogonal systems and their relation to other important formulas in the field as, e.g., to addition formulas and Laplace type integral representations. In particular, we will take the Jacobi polynomial systems as examples which have been extensively studied. While product and addition formulas for the ultraspherical polynomials are well known for quite a time, the respective formulas for the Jacobi polynomials were found by Koornwinder within the last decade. See, e.g., [18], [19], and, for the historical background, also the introductory survey paper of Koornwinder's thesis [17]. He gave several different proofs, some of which are group theoretic in nature, while others are analytic. It turns out that, for Jacobi polynomials, one of the four properties: Laplace representation, product formula, degenerate and ordinary addition formula implies the three others [19]. This may be a general feature, also for other orthogonal systems. Let us mention only one analytic proof of the product formula [18]. It also proceeds from a partial differential equation, but now of elliptic type, i.e., from the biaxially symmetric potential equation

$$\left(D^{\beta}_{0,x} + D^{\alpha}_{0,t}\right) F(x,t) = 0, \tag{1.9}$$

where the same differential operators appear as in the Hankel case (cf. (1.4)). Solving (1.9) in two different ways leads, after some transformations, to Bateman's bilinear sum which expresses the product of the Jacobi polynomials $P^{\alpha,\beta}_n(x) P^{\alpha,\beta}_n(t)$ as a linear combination of the terms $(x+t)^k P^{\alpha,\beta}_k((1+xt)/(x+t))$. An application to the Laplace type integral representations of the Jacobi polynomials (which can be derived from the corresponding formula for the ultraspherical polynomials by means of fractional integration) then yields the Jacobi product formula immediately.

2. Translation and Riemann function. We begin with a brief review of Riemann's method (cf. [27] and, e.g., [4], [10]). The Cauchy problem (1.5) can be reformulated so as to have homogeneous initial conditions. To this end we set $v(x,t)=u(x,t)-f(x)$, so that, with the notation $D^{\alpha}_{q,x,t}=D^{\alpha}_{q,x}-D^{\alpha}_{q,t}$,

$$\begin{aligned} &D^{\alpha}_{q,x,t}v(x,t)=-D^{\alpha}_{q,x,t}f(x) \qquad (0<t\leqq x),\\ &v(x,0)=0, \quad v_t(x,0)=0 \quad (x>0). \end{aligned} \tag{2.1}$$

Denote by $A_q(\xi,\tau;\ x,t)$ the Riemann function associated with the operator

$$D^{\alpha}_{q,\xi,\tau}=\frac{\partial^2}{\partial\xi^2}-\frac{\partial^2}{\partial\tau^2}+\frac{2\alpha+1}{\xi}\frac{\partial}{\partial\xi}-\frac{2\alpha+1}{\tau}\frac{\partial}{\partial\tau}-[q(\xi)-q(\tau)],$$

i.e., $A_q(\xi,\tau;x,t)=v^*(\xi,\tau;x,t)$ where v^* solves the characteristic boundary value problem

(2.2)

$$\begin{aligned} &\left(D^{\alpha}_{q,\xi,\tau}\right)^*v^*(\xi,\tau)\\ &\quad=v^*_{\xi\xi}-v^*_{\tau\tau}-\left(\frac{2\alpha+1}{\xi}v^*\right)_{\xi}+\left(\frac{2\alpha+1}{\tau}v^*\right)_{\tau}-[q(\xi)-q(\tau)]v^*=0\\ &\qquad\qquad \text{if } (\xi,\tau)\in\Delta_{xt}=\{(\xi,\tau);\ 0<\tau<t,\ x-t+\tau<\xi<x+t-\tau\},\\ &v^*_{\xi}+v^*_{\tau}=\left(\alpha+\frac{1}{2}\right)\left(\frac{1}{\xi}+\frac{1}{\tau}\right)v^* \quad \text{if } \xi-\tau=x-t,\\ &v^*_{\xi}-v^*_{\tau}=\left(\alpha+\frac{1}{2}\right)\left(\frac{1}{\xi}-\frac{1}{\tau}\right)v^* \quad \text{if } \xi+\tau=x+t,\\ &v^*(x,t;\ x,t)=1. \end{aligned}$$

Applying Green's theorem to

$$v^*D^{\alpha}_{q,\xi,\tau}v=v^*D^{\alpha}_{q,\xi,\tau}v-v\left(D^{\alpha}_{q,\xi,\tau}\right)^*v^*,$$

and choosing for the region of integration the triangle $\Delta_{xt,\varepsilon}$ with vertices $R(x,t)$ and $P_{\varepsilon}(x-t+\varepsilon,\varepsilon)$, $Q_{\varepsilon}(x+t-\varepsilon,\varepsilon)$ for some $\varepsilon>0$, one obtains

$$\begin{aligned} &\iint_{\Delta_{xt,\varepsilon}} v^*D^{\alpha}_{q,\xi,\tau}v\,d\xi\,d\tau\\ &\quad=\int_{\partial\Delta_{xt,\varepsilon}}\left[\left(v^*v_{\tau}-vv^*_{\tau}+\frac{2\alpha+1}{\tau}vv^*\right)d\xi+\left(v^*v_{\xi}-vv^*_{\xi}+\frac{2\alpha+1}{\xi}vv^*\right)d\tau\right]\\ &\quad=(vv^*)_{P_{\varepsilon}}+(vv^*)_{Q_{\varepsilon}}-2(vv^*)_R+\int_{x-t+\varepsilon}^{x+t-\varepsilon}\left(v^*v_{\tau}-vv^*_{\tau}+\frac{2\alpha+1}{\tau}vv^*\right)\Big|_{\tau=\varepsilon}d\xi. \end{aligned}$$

Here $(vv^*)_{P_{\varepsilon}}$ denotes the value of vv^* at the vertex P_{ε}, etc. Under the assumption that v^* and $((2\alpha+1)/\tau)v^*-v^*_{\tau}$ remain continuous on $\Delta_{xt,\varepsilon}$ for ε tending to $0+$, this leads in the limit to

$$v(x,t)=-\frac{1}{2}\iint_{\Delta_{xt}} v^*D^{\alpha}_{q,\xi,\tau}v\,d\xi\,d\tau \qquad \left(\alpha\geqq-\frac{1}{2}\right),$$

in view of the initial data of v. As will be shown in §§4 and 5, the two continuity conditions are satisfied if $q(x)=x^2$. In view of (2.1), one can replace $v(\xi,\tau)$ by $-f(\xi)$

in the last integral, so that a second application of Green's theorem yields

$$v(x,t)=\frac{1}{2}\iint_{\Delta_{xt}} v^* D^{\alpha}_{q,\xi,\tau} f\,d\xi\,d\tau$$

$$=\frac{1}{2}\{(fv^*)_{P_0}+(fv^*)_{Q_0}\}-(fv^*)_R+\frac{1}{2}\lim_{\varepsilon\to 0+}\int_{x-t+\varepsilon}^{x+t-\varepsilon}\left(\frac{2\alpha+1}{\tau}v^*-v^*_\tau\right)\Big|_{\tau=\varepsilon} f\,d\xi.$$

Since $(v^*)_R=1$, and $(v^*)_{P_0}=(v^*)_{Q_0}=0$ if $\alpha>-\frac{1}{2}$, and $=1$ if $\alpha=-\frac{1}{2}$ (cf. (4.12), (4.13)), one finally derives

$$u(x,t)=v(x,t)+f(x)$$

$$=\begin{cases}\displaystyle\int_{x-t}^{x+t} f(\xi)w^{\alpha}_q(x,t,\xi)\,d\xi, & \alpha>-\frac{1}{2},\\ \displaystyle\frac{1}{2}\{f(x-t)+f(x+t)\}+\int_{x-t}^{x+t} f(\xi)w^{\alpha}_q(x,t,\xi)\,d\xi, & \alpha=-\frac{1}{2},\end{cases}$$

for $0<t\leqq x$, where

$$w^{\alpha}_q(x,t,\xi)=\frac{1}{2}\lim_{\tau\to 0+}\left\{\frac{2\alpha+1}{\tau}A^{\alpha}_q(\xi,\tau;\,x,t)-\frac{\partial}{\partial\tau}A^{\alpha}_q(\xi,\tau;\,x,t)\right\} \tag{2.3}$$

$$=-\frac{1}{2}\lim_{\tau\to 0+}\left\{\tau^{2\alpha+1}\frac{\partial}{\partial\tau}\left[\tau^{-2\alpha-1}A^{\alpha}_q(\xi,\tau;\,x,t)\right]\right\}.$$

For $0<x<t$, set $u(x,t)=u(t,x)$. If $w^{\alpha}_q(x,t,\xi)$ is symmetric with respect to x and t for each $|x-t|<\xi<x+t$, the kernel of the corresponding generalized translation is then given by

$$K^{\alpha}_q(x,t,\xi)=\xi^{-2\alpha-1}w^{\alpha}_q(x,t,\xi). \tag{2.4}$$

This relation will be the basis of the proof of Theorem 1 in §5.

In view of the importance of Riemann's method [27] for various problems concerning partial differential equations, many authors have tried to determine as many Riemann functions as possible in closed form. For example, there are contributions by Chaundy, Cohn, Copson, Henrici, Vekua, and, particularly in recent years, by Bauer, Florian, Lanckau, Püngel, Wahlberg, and Wallner, cf. e.g. [21], [26], [31] and the literature cited there. Here, one essential step is to introduce appropriate auxiliary variables for representing the Riemann function, the number of which increases with the generality of the underlying PDE. For example, let us take the boundary value problem, which will be of particular interest in the following (cf. §4),

$$\begin{aligned} &V_{\zeta\eta}-\left[C_0+C_1(\zeta-\eta)^{-2}\right]V=0, \qquad C_0,\,C_1=\text{const.},\\ &V(\zeta,\eta_0;\,\zeta_0,\eta_0)=V(\zeta_0,\eta;\,\zeta_0,\eta_0)=1.\end{aligned} \tag{2.5}$$

Its solution follows at once from a result of Henrici [14] by changing the variable η into $-\eta$. With the two auxiliary variables

$$w_0=(\zeta-\zeta_0)(\eta-\eta_0), \qquad w=w_0\left[(\zeta-\eta)(\zeta_0-\eta_0)\right]^{-1} \tag{2.6a}$$

the solution is given as the confluent hypergeometric function

$$V(\zeta,\eta;\,\zeta_0,\eta_0)=\Xi_2(p,1-p;\,1;\,w,C_0w_0) \tag{2.6b}$$

$$=\sum_{n=0}^{\infty}\sum_{k=0}^{\infty}\frac{(p)_n(1-p)_n}{(n+k)!n!k!}w^n(C_0w_0)^k,$$

where $|w|<1$, w_0 is arbitrary, and $p(1-p)=C_1$. This can also be derived from more general results of Lanckau [21], Püngel [26], and Wallner [31]. Let us mention that Püngel gave his Riemann function in a somewhat different form, namely as a power series expansion in powers of w_0 with coefficients being polynomials of several variables. When applied to our example (2.5), his result reduces to

$$V(\zeta,\eta;\zeta_0,\eta_0)=\sum_{k=0}^{\infty} Q^k(v)\frac{(w_0)^k}{k!^2}, \qquad v=\frac{w}{w_0},$$

$$Q^k(v)=\sum_{n=0}^{k}\binom{k}{n}(p)_n(1-p)_n C_0^{k-n}v^n,$$

which coincides with (2.6b).

3. Auxiliary results. We collect some properties of the hypergeometric function to be used in the sequel. (See, e.g., [9], [29].) The transformations

(3.1a) $$F(a,b;\,c;\,z)=(1-z)^{-a}F\left(a,c-b;\,c;\,\frac{z}{z-1}\right)$$

(3.1b) $$F(a,b;\,c;\,z)=(1-z)^{-b}F\left(c-a,b;\,c;\,\frac{z}{z-1}\right) \qquad \left(|z|<1,\ \left|\frac{z}{z-1}\right|<1\right),$$

(3.1c) $$F(a,b;\,c;\,z)=(1-z)^{c-a-b}F(c-a,c-b;\,c;\,z) \qquad (|z|<1)$$

are due to Euler as is the integral representation

$$F(a,b;\,c;\,z)=\frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)}\int_0^1 t^{b-1}(1-t)^{c-b-1}(1-tz)^{-a}dt$$

for $\operatorname{Re}c>\operatorname{Re}b>0$, $|\arg(1-z)|<\pi$. The Gauss summation theorem reads

$$F(a,b;\,c;\,1)=\frac{\Gamma(c)\Gamma(c-a-b)}{\Gamma(c-a)\Gamma(c-b)} \qquad (c\neq 0,-1,\cdots;\ \operatorname{Re}c>\operatorname{Re}(a+b)).$$

Moreover, one has

$$\frac{d}{dz}F(a,b;\,c;\,z)=\frac{ab}{c}F(a+1,b+1;\,c+1;\,z),$$

$$F(a,b;\,c;\,z)=\frac{\Gamma(c)\Gamma(c-a-b)}{\Gamma(c-a)\Gamma(c-b)}F(a,b;\,a+b+1-c;\,1-z)$$

$$+\frac{\Gamma(c)\Gamma(a+b-c)}{\Gamma(c)\Gamma(b)}(1-z)^{c-a-b}F(c-a,c-b;\,c-a-b+1;\,1-z)$$

if $a+b<c<a+b+1$, $c\neq 0,-1,-2,\cdots$, and for $a,b\neq 0,-1,-2,\cdots$,

$$F(a,b;\,a+b;\,z)=\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\sum_{n=0}^{\infty}\frac{(a)_n(b)_n}{n!n!}[k_{n,0}-\log(1-z)](1-z)^n,$$

$$F(a,b;\,a+b+1;\,z)=\frac{\Gamma(a+b+1)}{\Gamma(a+1)\Gamma(b+1)}-(1-z)\frac{\Gamma(a+b+1)}{\Gamma(a)\Gamma(b)}\cdot$$

$$\sum_{n=0}^{\infty}\frac{(a+1)_n(b+1)_n}{(n+1)!n!}[k_{n,1}-\log(1-z)](1-z)^n,$$

where $k_{n,j}=\psi(n+1)+\psi(n+1+j)-\psi(a+n+j)-\psi(b+n+j)$ $(j=0,1)$. Hence the behavior of F for $z\to 1-$ can be described by

(3.2a)

$$|F(a,b;\ c;\ z)-F(a,b;\ c;\ 1)|$$

$$\leqq C\begin{cases}\left|\dfrac{ab\Gamma(c)\Gamma(c-a-b-1)}{\Gamma(c-a)\Gamma(c-b)}\right|(1-z) & \text{if } c>a+b+1,\\ \left|\dfrac{\Gamma(a+b+1)}{\Gamma(a)\Gamma(b)}\right|(1-z)\log\left(\dfrac{1}{1-z}\right) & \text{if } c=a+b+1,\\ \left|\dfrac{\Gamma(c)\Gamma(a+b-c)}{\Gamma(a)\Gamma(b)}\right|(1-z)^{c-a-b} & \text{if } a+b<c<a+b+1,\end{cases}$$

$$F(a,b;\ a+b;\ z)=\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\log\left(\frac{1}{1-z}\right)+O(1)\qquad(z\to 1-). \tag{3.2b}$$

The following lemma can be considered as a generalization of Euler's transformation.

LEMMA 1. *For $n\in\mathbb{P}$, $a<1$, $0\leqq\phi\leqq 1$, $\chi\geqq 0$, the function*

$$S(n,a;\ \phi,\chi)=\sum_{k=0}^{\infty}F(n+a,k;\ k+n+1;\ \phi)\frac{n!(-\chi)^k}{k!(k+n)!}$$

satisfies

$$|S(n,a;\ \phi,\chi)|\leqq\begin{cases}1, & a\leqq\dfrac{1}{2},\\ 1+\dfrac{\chi}{(1-a)(2-a)}, & \dfrac{1}{2}<a<1.\end{cases} \tag{3.3}$$

Moreover, if $a\leqq\frac{1}{2}$, $b>0$, $c>0$, $|z|<1$, and $|z/(z-1)|<1$, there holds

$$\sum_{n=0}^{\infty}\left\{\sum_{k=0}^{\infty}\frac{n!(-\chi)^k}{k!(k+n)!}\right\}\frac{(a)_n(b)_n}{(c)_n n!}z^n$$

$$=(1-z)^{-a}\sum_{n=0}^{\infty}S\left(n,a;\ \frac{z}{z-1},\chi\right)\frac{(a)_n(c-b)_n}{(c)_n n!}\left(\frac{z}{z-1}\right)^n. \tag{3.4}$$

Proof. Concerning (3.3), insert

$$F(n+a,k;\ k+n+1;\ 1)=\frac{(k+n)!\Gamma(1-a)}{n!\Gamma(k+1-a)}\qquad(a<1)$$

to deduce $S(n,a;\ 1,\chi)=j_{-a}(2\sqrt{\chi})$, from which the assertion for $\phi=1$ follows by means of the estimate of the Bessel functions [33, 3.31(1), (2)]

$$|j_\nu(x)|\leqq\begin{cases}1, & \nu\geqq-\dfrac{1}{2},\\ 1+\dfrac{x^2}{4(\nu+1)(\nu+2)}, & -1<\nu<-\dfrac{1}{2}\end{cases}\qquad(x\geqq 0). \tag{3.5}$$

Now let $0 \leqq \phi < 1$. In view of Euler's integral representation one has

$$S(n,a;\ \phi,\chi) = 1 + \int_0^1 \left\{ \sum_{k=1}^{\infty} \frac{(-\chi)^k t^{k-1}}{k!(k-1)!} \right\} g_{n,\phi}(t)\, dt,$$

where $g_{n,\phi}(t) = (1-t)^n (1-t\phi)^{-n-a}$. Interpreting the inner sum as the derivative of j_0, summation by parts gives

$$S(n,a;\ \phi,\chi) = j_0(2\sqrt{\chi})(1-\phi)^{-a}\delta_{n,0} - \int_0^1 j_0(2\sqrt{\chi t})\, g'_{n,\phi}(t)\, dt,$$

$\delta_{n,0}$ being Kronecker's symbol. In case $a \leqq 0$,

$$g'_{n,\phi}(t) = -[n(1-\phi) - a\phi(1-t)](1-t)^{n-1}(1-t\phi)^{-n-a-1}$$

is nonpositive on $0 \leqq t \leqq 1$ for each $n \in \mathbb{P}$, $0 \leqq \phi < 1$, so that, in view of (3.5),

$$|S(n,a;\ \phi,\chi)| \leqq (1-\phi)^{-a}\delta_{n,0} - \int_0^1 g'_{n,\phi}(t)\, dt \equiv 1 \qquad (\chi \geqq 0).$$

In the remaining cases $0 < a < 1$ we use

$$\begin{aligned} F(n+a,k;\ k+n+1;\ \phi) &= (1-\phi)^{1-a} F(n+1,k-a+1;\ k+n+1;\ \phi) \\ &= (1-\phi)^{1-a} \frac{\Gamma(k+n+1)}{\Gamma(k-a+1)\Gamma(n+a)} \int_0^1 t^{k-a}(1-t)^{n+a-1}(1-t\phi)^{-n-1} dt \end{aligned}$$

to deduce

$$S(n,a;\ \phi,\chi) = (1-\phi)^{1-a} \frac{\Gamma(n+1)}{\Gamma(n+a)\Gamma(1-a)} \cdot \int_0^1 j_{-a}(2\sqrt{\chi t})\, t^{-a}(1-t)^{n+a-1}(1-t\phi)^{-n-1} dt. \tag{3.6}$$

Applying (3.5) once more, one has

$$|S(n,a;\ \phi,\chi)| \leqq (1-\phi)^{1-a} F(n+1,1-a;\ n+1;\ \phi) \equiv 1$$

for $0 < a \leqq \frac{1}{2}$ and, for $\frac{1}{2} < a < 1$,

$$\begin{aligned} |S(n,a;\ \phi,\chi)| &\leqq 1 + (1-\phi)^{1-a} \frac{\chi}{(n+1)(2-a)} F(n+1,2-a;\ n+2;\ \phi) \\ &\leqq 1 + \frac{\chi}{(1-a)(2-a)}. \end{aligned}$$

This completes the proof of (3.3). To verify (3.4), we generalize a proof of Euler's transformation (3.1a) as given in [29, §1.7.1]. Setting

$$\frac{(a)_n (b)_n}{(c)_n n!} = \frac{(a)_n}{n!} F(-n,c-b;\ c;\ 1) = \sum_{j=0}^{n} \frac{(a)_j (c-b)_j (j+a)_{n-j} (-1)^j}{(c)_j j!(n-j)!}$$

for $b>0$, $c>0$, $n\in\mathbb{P}$, the left-hand side of (3.4) becomes

$$\sum_{n=0}^{\infty}\left\{\sum_{k=0}^{\infty}\frac{n!(-\chi)^k}{k!(k+n)!}\right\}\left\{\sum_{j=0}^{n}\frac{(a)_j(c-b)_j(j+a)_{n-j}(-1)^j}{(c)_j j!(n-j)!}\right\}z^n$$

$$=\sum_{j=0}^{\infty}\left\{\sum_{n=j}^{\infty}\left[\sum_{k=0}^{\infty}\frac{n!(-\chi)^k}{k!(k+n)!}\right]\frac{(j+a)_{n-j}}{(n-j)!}z^{n-j}\right\}\frac{(a)_j(c-b)_j}{(c)_j j!}(-z)^j.$$

By an index transformation, an interchange of summations and an application of (3.1a), the term in curly brackets can be written as

$$\sum_{k=0}^{\infty}\left[\sum_{n=0}^{\infty}\frac{(j+a)_n(n+j)!}{(n+j+k)!n!}z^n\right]\frac{(-\chi)^k}{k!}$$

$$=\sum_{k=0}^{\infty}F(j+a,j+1;\ k+j+1;\ z)\frac{j!(-\chi)^k}{(k+j)!k!}$$

$$=(1-z)^{-a-j}\sum_{k=0}^{\infty}F\left(j+a,k;\ k+j+1;\ \frac{z}{z-1}\right)\frac{j!(-\chi)^k}{(k+j)!k!},$$

which proves (3.4). □

4. The Riemann function associated with the Laguerre translation. In the characteristic boundary value problem (2.2) we choose the Laguerre potential $q(\xi)=\xi^2$ and substitute

$$\begin{aligned}\zeta(\xi,\tau)&=\frac{1}{4}(\xi+\tau)^2, \qquad \zeta_0=\zeta(x,t),\\ \eta(\xi,\tau)&=\frac{1}{4}(\xi-\tau)^2, \qquad \eta_0=\eta(x,t).\end{aligned}\tag{4.1}$$

Dropping the asterisk, we obtain

$$\begin{aligned}&v_{\zeta\eta}+\left(\frac{\alpha+1/2}{\zeta-\eta}v\right)_{\zeta}-\left(\frac{\alpha+1/2}{\zeta-\eta}v\right)_{\eta}-v=0 \quad \text{if } (\zeta,\eta)\in\Delta_{\zeta_0\eta_0},\\ &v_{\zeta}=\frac{\alpha+1/2}{\zeta-\eta}v \quad \text{if } \eta=\eta_0,\\ &v_{\eta}=-\frac{\alpha+1/2}{\zeta-\eta}v \quad \text{if } \zeta=\zeta_0,\\ &v(\zeta_0,\eta_0;\ \zeta_0,\eta_0)=1\end{aligned}\tag{4.2}$$

for $\alpha\geqq-\frac{1}{2}$, where now

$$\Delta_{\zeta_0\eta_0}=\{(\zeta,\eta);\ \eta_0<\eta<\zeta_0,\ \eta<\zeta<\zeta_0\}.$$

The boundary conditions in (4.2) can also be written as

$$v(\zeta,\eta;\ \zeta_0,\eta_0)=\left(\frac{\zeta-\eta}{\zeta_0-\eta_0}\right)^{\alpha+1/2} \quad \text{if } \eta=\eta_0 \text{ or } \zeta=\zeta_0.\tag{4.3}$$

We further use the transformation

$$v(\zeta,\eta;\ \zeta_0,\eta_0)=\left(\frac{\zeta-\eta}{\zeta_0-\eta_0}\right)^{\alpha+1/2}V(\zeta,\eta;\ \zeta_0,\eta_0)\tag{4.4}$$

to turn (4.2) into its selfadjoint form

$$V_{\zeta\eta}+\left[\frac{\alpha^2-1/4}{(\zeta-\eta)^2}-1\right]V=0 \quad \text{if } (\zeta,\eta)\in\Delta_{\zeta_0\eta_0}, \tag{4.5}$$
$$V(\zeta,\eta;\,\zeta_0,\eta_0)=1 \quad \text{if } \eta=\eta_0 \text{ or } \zeta=\zeta_0.$$

As a pattern for solving this problem we consider the corresponding boundary value problem for the Riemann function of the Hankel case

$$(V_H^\alpha)_{\zeta\eta}+\frac{\alpha^2-1/4}{(\zeta-\eta)^2}V_H^\alpha=0 \quad \text{if } (\zeta,\eta)\in\Delta_{\zeta_0\eta_0},$$
$$V_H^\alpha(\zeta,\eta;\,\zeta_0,\eta_0)=1 \quad \text{if } \eta=\eta_0 \text{ or } \zeta=\zeta_0.$$

(Throughout the paper the subscripts H and L stand for Hankel and Laguerre, respectively.) Its solution can be represented using one auxiliary variable only:

$$V_H^\alpha(\zeta,\eta;\,\zeta_0,\eta_0)=F\left(\frac{1}{2}+\alpha,\frac{1}{2}-\alpha;\,1;\,w\right), \qquad w=\frac{(\zeta-\zeta_0)(\eta-\eta_0)}{(\zeta-\eta)(\zeta_0-\eta_0)}.$$

But while the hypergeometric function F converges in the unit disk only, the argument $w(\zeta,\eta;\,\zeta_0,\eta_0)$ ranges over the whole negative semiaxis when (ζ,η) ranges over $\Delta_{\zeta_0\eta_0}$. To establish the complete solution by analytic continuation one therefore has to use Euler's transformations (3.1a) and (3.1b) according to $\alpha\leqq 0$ or $\alpha>0$. Hence, setting

$$\chi=(\zeta_0-\zeta)(\eta-\eta_0), \quad \psi=(\zeta_0-\eta)(\zeta-\eta_0), \quad \phi=\frac{\chi}{\psi}=\frac{w}{w-1} \tag{4.6}$$

and returning to v via (4.4) again, one obtains

$$v_H^\alpha(\zeta,\eta;\,\zeta_0,\eta_0) \tag{4.7}$$
$$=\begin{cases}\dfrac{\zeta-\eta}{(\zeta_0-\eta_0)^{2\alpha}}\psi^{\alpha-1/2}F\left(\dfrac{1}{2}-\alpha,\dfrac{1}{2}-\alpha;\,1;\,\phi\right), & \alpha>0,\\[2ex] (\zeta-\eta)^{2\alpha+1}\psi^{-\alpha-1/2}F\left(\dfrac{1}{2}+\alpha,\dfrac{1}{2}+\alpha;\,1;\,\phi\right), & -\dfrac{1}{2}\leqq\alpha\leqq 0.\end{cases}$$

Notice that ϕ vanishes on the characteristics $\eta=\eta_0$ and $\zeta=\zeta_0$, in accordance with (4.3). If $(\zeta,\eta)\in\Delta_{\zeta_0\eta_0}$ tends to the initial line $\zeta=\eta$, ϕ tends to $1-$, and the asymptotic behavior of the hypergeometric function implies that the Riemann function v_H^α is still continuous on $\eta_0<\zeta=\eta<\zeta_0$ with

$$v_H^\alpha(\zeta,\zeta;\,\zeta_0,\eta_0)=\begin{cases}0, & \alpha>-\dfrac{1}{2},\\[2ex] 1, & \alpha=-\dfrac{1}{2}.\end{cases} \tag{4.8}$$

When substituting (4.1) into (4.7) and (4.8), the Riemann function of the Hankel case in the sense of (2.2) is finally given for $(\xi,\tau)\in\overline{\Delta_{xt}}$ by

$$A_H^\alpha(\xi,\tau;\,x,t)=v_H^\alpha\left(\frac{1}{4}(\xi+\tau)^2,\frac{1}{4}(\xi-\tau)^2;\,\frac{1}{4}(x+t)^2,\frac{1}{4}(x-t)^2\right). \tag{4.9}$$

Returning to problem (4.5), we use the solution (2.6) of (2.5) for the case $C_0=1$ and $C_1=\frac{1}{4}-\alpha^2$ to get

$$V_L^\alpha(\zeta,\eta;\,\zeta_0,\eta_0)=\sum_{n=0}^{\infty}\left\{\sum_{k=0}^{\infty}\frac{n!(-\chi)^k}{k!(k+n)!}\right\}\frac{(1/2+\alpha)_n(1/2-\alpha)_n}{n!n!}w^n$$

$$=\sum_{n=0}^{\infty}j_n(2\sqrt{\chi})\frac{(1/2+\alpha)_n(1/2-\alpha)_n}{n!n!}w^n$$

for those ζ, η for which the series converges absolutely, i.e., for $|w|<1$. In the particular cases $\alpha=-\frac{1}{2}$ and $\alpha=\frac{1}{2}$, the result reduces to

$$V_L^{\pm 1/2}(\zeta,\eta;\,\zeta_0,\eta_0)=j_0(2\sqrt{\chi})=J_0\left(2\sqrt{(\zeta_0-\zeta)(\eta-\eta_0)}\right) \tag{4.10}$$

which is known as the Riemann function of the telegraph equation $V_{\zeta\eta}-V=0$, cf., e.g. [10, p. 133].

In order to obtain the solution for each $w\leqq 0$ by analytic continuation, we now have to apply (3.4) of Lemma 1. Choosing $a=\frac{1}{2}+\alpha$, $b=\frac{1}{2}-\alpha$ if $-\frac{1}{2}\leqq\alpha\leqq 0$, and $a=\frac{1}{2}-\alpha$, $b=\frac{1}{2}+\alpha$ if $\alpha>0$, as well as $c=1$, and inverting the transformation (4.4), it follows for $(\zeta,\eta)\in\Delta_{\zeta_0\eta_0}$ that

(4.11)

$$v_L^\alpha(\zeta,\eta;\,\zeta_0,\eta_0)$$

$$=\begin{cases}\dfrac{\zeta-\eta}{(\zeta_0-\eta_0)^{2\alpha}}\psi^{\alpha-1/2}\displaystyle\sum_{n=0}^{\infty}S\left(n,\frac{1}{2}-\alpha;\,\phi,\chi\right)\frac{(1/2-\alpha)_n(1/2-\alpha)_n}{n!n!}\phi^n, & \alpha>0,\\[2ex] (\zeta-\eta)^{2\alpha+1}\psi^{-\alpha-1/2}\displaystyle\sum_{n=0}^{\infty}S\left(n,\frac{1}{2}+\alpha;\,\phi,\chi\right)\frac{(1/2+\alpha)_n(1/2+\alpha)_n}{n!n!}\phi^n, & -\frac{1}{2}\leqq\alpha\leqq 0,\end{cases}$$

where

$$S\left(n,\frac{1}{2}\pm\alpha;\,\phi,\chi\right)=\sum_{k=0}^{\infty}F\left(n+\frac{1}{2}\pm\alpha,k;\,k+n+1;\,\phi\right)\frac{n!(-\chi)^k}{k!(k+n)!}.$$

Notice that $v_H^\alpha(\zeta,\eta;\,\zeta_0,\eta_0)$ in (4.7) is of the same form as (4.11) except for the factors $S(n,\frac{1}{2}\pm\alpha;\,\phi,\chi)$. As in the Hankel case, v_L^α depends continuously on the boundary conditions (4.3). Since $S(n,1/2-|\alpha|,\phi,\chi)$ tends to $j_{|\alpha|-1/2}(2\sqrt{\chi})$ for each $n\in\mathbb{P}$ if ϕ tends to $1-$, v_L^α has a continuous extension to the initial line $\eta_0<\zeta=\eta<\zeta_0$, and

$$v_L^\alpha(\zeta,\zeta;\,\zeta_0,\eta_0)=\begin{cases}0, & \alpha>-\dfrac{1}{2},\\[2ex] J_0\left(2\sqrt{(\zeta_0-\zeta)(\zeta-\eta_0)}\right), & \alpha=-\dfrac{1}{2}.\end{cases} \tag{4.12}$$

After substituting (4.1) into (4.11) and (4.12) one obtains the final version of the Riemann function in the Laguerre case,

$$A_L^\alpha(\xi,\tau;\,x,t)=v_L^\alpha\left(\frac{1}{4}(\xi+\tau)^2,\,\frac{1}{4}(\xi-\tau)^2;\,\frac{1}{4}(x+t)^2,\,\frac{1}{4}(x-t)^2\right), \tag{4.13}$$

when $(\xi,\tau)\in\overline{\Delta_{xt}}$.

5. The kernel of the Laguerre translation, proof of Theorem 1. In view of (2.3) and (2.4), the kernel of the Laguerre translation can be derived from the representation of the Riemann function by

$$
\begin{aligned}
&w_L^\alpha(x,t,\xi) = -\frac{1}{2}\lim_{\tau\to 0+} G_L^\alpha(\tau),\\
&G_L^\alpha(\tau) = G_L^\alpha(\tau;\, x,t,\xi) = \tau^{2\alpha+1}\frac{\partial}{\partial\tau}\left[\tau^{-2\alpha-1}A_L^\alpha(\xi,\tau;\, x,t)\right].
\end{aligned}
\tag{5.1}
$$

The main purpose of this paragraph is to prove somewhat more than needed for the proof of Theorem 1, namely an asymptotic expansion of $G_L^\alpha(\tau)$ for $\tau\to 0+$ (see (5.11)).

Again, let us first consider the Hankel case. Setting

$$
\begin{aligned}
&\phi(\xi,\tau;\, x,t) = \chi(\xi,\tau;\, x,t)/\psi(\xi,\tau;\, x,t),\\
&\chi(\xi,\tau;\, x,t) = \frac{1}{16}\left[(x+t)^2-(\xi+\tau)^2\right]\left[(\xi-\tau)^2-(x-t)^2\right],\\
&\psi(\xi,\tau;\, x,t) = \frac{1}{16}\left[(x+t)^2-(\xi-\tau)^2\right]\left[(\xi+\tau)^2-(x-t)^2\right],
\end{aligned}
$$

(cf. (4.1), (4.6)) and writing the Riemann function (4.9) in the form

$$
\begin{aligned}
&A_H^\alpha(\xi,\tau;\, x,t) = a_H^{|\alpha|}(\xi,\tau;\, x,t)\cdot\begin{cases}\xi\tau(xt)^{-2\alpha}, & \alpha>0,\\ (\xi\tau)^{2\alpha+1}, & -\frac{1}{2}\leqq\alpha\leqq 0,\end{cases}\\
&a_H^\gamma(\xi,\tau;\, x,t) = \psi^{\gamma-1/2}F\left(\frac{1}{2}-\gamma,\frac{1}{2}-\gamma;\,1;\,\phi\right)\qquad(\gamma\geqq 0),
\end{aligned}
\tag{5.2}
$$

the kernel of the Hankel translation is given by

$$
\begin{aligned}
&w_H^\alpha(x,t,\xi) = -\frac{1}{2}\lim_{\tau\to 0+} G_H^\alpha(\tau),\\
&G_H^\alpha(\tau) = \tau^{2\alpha+1}\frac{\partial}{\partial\tau}\left[\tau^{-2\alpha-1}A_H^\alpha(\xi,\tau;\, x,t)\right]\\
&\qquad = \begin{cases}\xi(xt)^{-2\alpha}\left\{-2\alpha\, a_H^\alpha(\tau)+\tau\frac{\partial}{\partial\tau}a_H^\alpha(\tau)\right\}, & \alpha>0,\\ (\xi\tau)^{2\alpha+1}\frac{\partial}{\partial\tau}a_H^{-\alpha}(\tau), & -\frac{1}{2}\leqq\alpha\leqq 0.\end{cases}
\end{aligned}
$$

For $\alpha=-\frac{1}{2}$ and $\alpha=\frac{1}{2}$ one immediately obtains

$$
G_H^{-1/2}(\tau)=0,\qquad G_H^{1/2}(\tau)=-\xi(xt)^{-1}.
$$

In the remaining cases, the asymptotic expansion of $G_H^\alpha(\tau)$ for $\tau\to 0+$ can essentially be deduced from

LEMMA 2. *Let* $\gamma\geqq 0$ *and let* $\psi_0=\psi(\xi,0;\, x,t)$. *Then*

$$
\text{(i)}\qquad a_H^\gamma(\tau) = \frac{\Gamma(2\gamma)}{\Gamma^2(\gamma+1/2)}\psi_0^{\gamma-1/2}+\begin{cases}O(\tau^{2\gamma}), & 0<\gamma<\frac{1}{2},\\ O(\tau), & \gamma>\frac{1}{2}\ (\tau\to 0+),\end{cases}
$$

(ii)

$$\frac{\partial}{\partial\tau}a_H^{\gamma}(\tau)=\begin{cases}-\dfrac{\Gamma(1-2\gamma)}{\Gamma^2(1/2-\gamma)}\psi_0^{-\gamma-1/2}(\xi xt)^{2\gamma}\tau^{2\gamma-1}+\mathcal{O}(\tau^{2\gamma}), & 0\leqq\gamma\leqq\dfrac{1}{2},\\ O(\tau^{2\gamma-1}), & \dfrac{1}{2}<\gamma<1,\\ O\left(\tau\log\dfrac{1}{\tau}\right), & \gamma=1,\\ O(\tau), & \gamma>1 \qquad (\tau\to 0+).\end{cases}$$

Proof. By definition, one has

$$1-\phi=\psi^{-1}\xi\tau xt, \qquad \phi'(\tau)=\psi^{-1}\left[\chi'(\tau)-\phi\psi'(\tau)\right],$$
$$\chi'(\tau)=-\frac{1}{2}\xi xt+\frac{\tau}{4}(\xi^2+x^2+t^2-\tau^2), \qquad \psi'(\tau)=\frac{1}{2}\xi xt+\frac{\tau}{4}(\xi^2+x^2+t^2-\tau^2).$$

The differentiability of ψ and (3.2) immediately implies the assertion of part (i). Concerning part (ii), an application of

$$\begin{aligned}&F\left(\frac{1}{2}-\gamma,\frac{1}{2}-\gamma;1;\phi\right)+\left(\frac{1}{2}-\gamma\right)F\left(\frac{3}{2}-\gamma,\frac{3}{2}-\gamma;2;\phi\right)\cdot\phi\\ &\qquad=F\left(\frac{3}{2}-\gamma,\frac{1}{2}-\gamma;1;\phi\right),\\ &F\left(\frac{3}{2}-\gamma,\frac{1}{2}-\gamma;1;\phi\right)+\left(\frac{1}{2}-\gamma\right)F\left(\frac{3}{2}-\gamma,\frac{3}{2}-\gamma;2;\phi\right)\\ &\qquad=\left(\frac{3}{2}-\gamma\right)F\left(\frac{3}{2}-\gamma,\frac{5}{2}-\gamma;2;\phi\right)(1-\phi),\\ &F\left(\frac{3}{2}-\gamma,\frac{1}{2}-\gamma;1;\phi\right)-\left(\frac{1}{2}-\gamma\right)F\left(\frac{3}{2}-\gamma,\frac{3}{2}-\gamma;2;\phi\right)\\ &\qquad=\left(\frac{1}{2}+\gamma\right)F\left(\frac{3}{2}-\gamma,\frac{1}{2}-\gamma;2;\phi\right)\end{aligned}$$

for $\phi<1$ leads to

$$\begin{aligned}&\frac{\partial}{\partial\tau}a_H^{\gamma}(\tau)\\ &\quad=\left(\gamma-\frac{1}{2}\right)\psi^{\gamma-3/2}\psi'(\tau)F\left(\frac{1}{2}-\gamma,\frac{1}{2}-\gamma;1;\phi\right)\\ &\qquad\qquad+\psi^{\gamma-1/2}\left(\frac{1}{2}-\gamma\right)^2F\left(\frac{3}{2}-\gamma,\frac{3}{2}-\gamma;2;\phi\right)\phi'(\tau)\\ &\quad=\left(\gamma-\frac{1}{2}\right)\psi^{\gamma-3/2}\Big\{\psi'(\tau)F\left(\frac{3}{2}-\gamma,\frac{1}{2}-\gamma;1;\phi\right)\\ &\qquad\qquad-\left(\frac{1}{2}-\gamma\right)\chi'(\tau)F\left(\frac{3}{2}-\gamma,\frac{3}{2}-\gamma;2;\phi\right)\Big\}\\ &\quad=\left(\gamma-\frac{1}{2}\right)\psi^{\gamma-3/2}\Big\{\frac{1}{2}\xi xt\left(\frac{3}{2}-\gamma\right)F\left(\frac{3}{2}-\gamma,\frac{5}{2}-\gamma;2;\phi\right)(1-\phi)\\ &\qquad\qquad+\frac{\tau}{4}(\xi^2+x^2+t^2-\tau^2)\left(\frac{1}{2}+\gamma\right)F\left(\frac{3}{2}-\gamma,\frac{1}{2}-\gamma;2;\phi\right)\Big\}.\end{aligned}$$

The assertion then follows by (3.1c) and (3.2). For $0 \leqq \gamma < \frac{1}{2}$, in particular, the first term in curly brackets gives

$$\frac{1}{2}\xi xt\left(\frac{3}{2}-\gamma\right)\frac{\Gamma(2-2\gamma)}{\Gamma(3/2-\gamma)\Gamma(5/2-\gamma)}(1-\phi)^{2\gamma-1}+O\big((1-\phi)^{2\gamma}\big) \qquad (\phi\to 1-)$$

$$=\frac{\Gamma(1-2\gamma)}{\Gamma(1/2-\gamma)\Gamma(3/2-\gamma)}\psi_0^{1-2\gamma}(\xi xt)^{2\gamma}\tau^{2\gamma-1}+O(\tau^{2\gamma}) \qquad (\tau\to 0+),$$

while the second term can be neglected. □

An application of Lemma 2 now yields

$$(5.3)\quad G_H^\alpha(\tau)=-\frac{\Gamma(2\alpha+1)}{\Gamma^2(\alpha+1/2)}\xi(xt)^{-2\alpha}\left(\frac{1}{16}\big[(x+t)^2-\xi^2\big]\big[\xi^2-(x-t)^2\big]\right)^{\alpha-1/2}$$

$$+\begin{cases} O(\tau) & \text{if } \alpha>\frac{1}{2} \text{ or } -\frac{1}{2}<\alpha\leqq 0\\ O(\tau^{2\alpha}) & \text{if } 0<\alpha<\frac{1}{2}\end{cases} \qquad (\tau\to 0+).$$

In the Laguerre case, we write the Riemann function (4.13) in a form analogous to (5.2),

$$A_L^\alpha(\xi,\tau;\,x,t)=a_L^{|\alpha|}(\xi,\tau;\,x,t)\cdot\begin{cases}\xi\tau(xt)^{-2\alpha}, & \alpha>0,\\ (\xi\tau)^{2\alpha+1}, & -\frac{1}{2}\leqq\alpha\leqq 0,\end{cases}$$

$$a_L^\gamma(\xi,\tau;\,x,t)=\psi^{\gamma-1/2}\sum_{n=0}^{\infty}S\left(n,\frac{1}{2}-\gamma;\,\phi,\chi\right)\frac{(1/2-\gamma)_n(1/2-\gamma)_n}{n!n!}\phi^n \qquad (\gamma\geqq 0).$$

Thus, $G_L^\alpha(\tau)$ in (5.1) can be expressed by

$$(5.4)\qquad G_L^\alpha(\tau)=\begin{cases}\xi(xt)^{-2\alpha}\left\{-2\alpha\, a_L^\alpha(\tau)+\tau\frac{\partial}{\partial\tau}a_L^\alpha(\tau)\right\}, & \alpha>0,\\ (\xi\tau)^{2\alpha+1}\frac{\partial}{\partial\tau}a_L^{-\alpha}(\tau), & -\frac{1}{2}\leqq\alpha\leqq 0.\end{cases}$$

When $\alpha=-\frac{1}{2}$ or $\alpha=\frac{1}{2}$, use $a_L^{1/2}(\tau)=j_0(2\sqrt{\chi})$ to get

$$(5.5)\qquad \begin{aligned} G_L^{-1/2}(\tau)&=-j_1\big(2\sqrt{\chi}\big)\chi'(\tau),\\ G_L^{1/2}(\tau)&=-\xi(xt)^{-1}\big\{j_0\big(2\sqrt{\chi}\big)+\tau j_1\big(2\sqrt{\chi}\big)\chi'(\tau)\big\}.\end{aligned}$$

Otherwise, the following analogue of Lemma 2 will be needed.

LEMMA 3. *Let* $\gamma\geqq 0$ *and let* $\psi_0=\chi_0=\psi(\xi,0;\,x,t)$. *Then*

(i)

$$a_L^\gamma(\tau)=\frac{\Gamma(2\gamma)}{\Gamma^2(\gamma+1/2)}\psi_0^{\gamma-1/2}j_{\gamma-1/2}\big(2\sqrt{\chi_0}\big)+\begin{cases}O(\tau^{2\gamma}), & 0<\gamma<\frac{1}{2},\\ O(\tau), & \gamma>\frac{1}{2}\end{cases} \qquad (\tau\to 0+),$$

(ii)

$$\frac{\partial}{\partial\tau}a_L^{\gamma}(\tau)=\begin{cases}-\dfrac{\Gamma(1-2\gamma)}{\Gamma^2(1/2-\gamma)}\dfrac{(\xi xt)^{2\gamma}}{\psi_0^{\gamma+1/2}}j_{-\gamma-1/2}\left(2\sqrt{\chi_0}\right)\tau^{2\gamma-1}+\begin{cases}O\left(\log\dfrac{1}{\tau}\right), & \gamma=0,\\ O(1), & 0<\gamma<\dfrac{1}{2},\end{cases}\\ O(1), \qquad \gamma>\dfrac{1}{2}\end{cases}\qquad(\tau\to 0+).$$

Proof. First we consider the cases $\gamma>\frac{1}{2}$ of part (i). By the mean value theorem, for each $n,k\in\mathbb{P}$ there exist $\tilde{\phi}\in(\phi,1)$ such that

$$\begin{aligned}F\left(n+\frac{1}{2}-\gamma,k;\ n+k+1;\ \phi\right)&=F\left(n+\frac{1}{2}-\gamma,k;\ n+k+1;\ 1\right)-\frac{(n+1/2-\gamma)k}{n+k+1}\\&\quad\cdot F\left(n+\frac{3}{2}-\gamma,k+1;\ n+k+2;\ \tilde{\phi}\right)(1-\phi)\\&=\frac{\Gamma(n+k+1)\Gamma(\gamma+1/2)}{\Gamma(n+1)\Gamma(k+\gamma+1/2)}-\frac{(n+1/2-\gamma)k\Gamma(n+k+1)}{\Gamma(k+1)\Gamma(n+1)}\\&\quad\cdot\int_0^1 t^k(1-t)^n(1-t\tilde{\phi})^{-n-3/2+\gamma}dt(1-\phi).\end{aligned}$$

So one has

$$\begin{aligned}&\left|S\left(n,\frac{1}{2}-\gamma;\ \phi,\chi\right)-j_{\gamma-1/2}(2\sqrt{\chi})\right|\\&=\left|\left(n+\frac{1}{2}-\gamma\right)\int_0^1\sum_{k=1}^{\infty}\frac{(-\chi t)^k}{(k-1)!k!}(1-t)^n(1-t\tilde{\phi})^{-n-3/2+\gamma}dt(1-\phi)\right|\\&\leqq\left|n+\frac{1}{2}-\gamma\right|\chi\int_0^1\left|j_1(2\sqrt{\chi t})\right|t\left(\frac{1-t}{1-t\tilde{\phi}}\right)^n(1-t\tilde{\phi})^{\gamma-3/2}dt(1-\phi)\\&\leqq C\left|n+\frac{1}{2}-\gamma\right|\chi(1-\phi)\end{aligned}$$

with a constant C depending only on γ, and consequently

$$\begin{aligned}&\left|a_L^{\gamma}(\tau)-\psi^{\gamma-1/2}F\left(\frac{1}{2}-\gamma,\frac{1}{2}-\gamma;1;\phi\right)j_{\gamma-1/2}(2\sqrt{\chi})\right|\\&\qquad\leqq C\psi^{\gamma-1/2}\chi\left|F\left(\frac{1}{2}-\gamma,\frac{3}{2}-\gamma;1;\phi\right)\right|(1-\phi)\leqq C(1-\phi).\end{aligned}$$

For $0<\gamma<\frac{1}{2}$, we use the representation (3.6) of $S(n,a;\ \phi,\chi)$ with $a=\frac{1}{2}-\gamma$ and the identity

$$\sum_{n=0}^{\infty}\frac{(1/2-\gamma)_n}{n!}\left(\frac{[1-t]\phi}{1-t\phi}\right)^n=\left(1-\frac{[1-t]\phi}{1-t\phi}\right)^{\gamma-1/2}=\left(\frac{1-\phi}{1-t\phi}\right)^{\gamma-1/2}\tag{5.6}$$

to obtain

$$\sum_{n=0}^{\infty} S\Big(n, \frac{1}{2}-\gamma;\ \phi, \chi\Big)\frac{(1/2-\gamma)_n(1/2-\gamma)_n}{n!n!}\phi^n$$

$$= \frac{(1-\phi)^{\gamma+1/2}}{\Gamma(1/2-\gamma)\Gamma(1/2+\gamma)}\int_0^1 j_{\gamma-1/2}\big(2\sqrt{\chi t}\big)t^{\gamma-1/2}(1-t)^{-1/2-\gamma}(1-t\phi)^{-1}$$

$$\cdot \sum_{n=0}^{\infty}\frac{(1/2-\gamma)_n}{n!}\left(\frac{[1-t]\phi}{1-t\phi}\right)^n dt$$

$$= \frac{(1-\phi)^{2\gamma}}{\Gamma(1/2-\gamma)\Gamma(1/2+\gamma)}\int_0^1 j_{\gamma-1/2}\big(2\sqrt{\chi t}\big)t^{\gamma-1/2}(1-t)^{-1/2-\gamma}(1-t\phi)^{-1/2-\gamma}dt.$$

Again by the mean value theorem, for each $0 \leqq t \leqq 1$ there exists $s(t) \in (t,1)$ such that

$$j_{\gamma-1/2}\big(2\sqrt{\chi t}\big) = j_{\gamma-1/2}\big(2\sqrt{\chi}\big) + \frac{\chi}{\gamma+1/2}j_{\gamma+1/2}\big(2\sqrt{\chi s(t)}\big)(1-t).$$

By (3.5) this gives

$$\left|a_L^{\gamma}(\tau) - \psi^{\gamma-1/2}(1-\phi)^{2\gamma}F\Big(\frac{1}{2}+\gamma, \frac{1}{2}+\gamma; 1;\ \phi\Big)j_{\gamma-1/2}\big(2\sqrt{\chi}\big)\right|$$

$$\leqq \psi^{\gamma-1/2}\chi\frac{(1-\phi)^{2\gamma}}{\Gamma(1/2-\gamma)\Gamma(3/2+\gamma)}$$

$$\cdot \int_0^1 \left|j_{\gamma+1/2}\big(2\sqrt{\chi s(t)}\big)\right| t^{\gamma-1/2}(1-t)^{1/2-\gamma}(1-t\phi)^{-1/2-\gamma}dt$$

$$\leqq \psi^{\gamma-1/2}\chi\frac{1/2-\gamma}{1/2+\gamma}(1-\phi)^{2\gamma}F\Big(\frac{1}{2}+\gamma, \frac{1}{2}+\gamma; 2;\ \phi\Big)$$

$$\leqq C(1-\phi)^{2\gamma}.$$

In view of the asymptotic expansions

$$\psi^{\gamma-1/2} = \psi_0^{\gamma-1/2} + O(\tau), \qquad j_{\gamma-1/2}\big(2\sqrt{\chi}\big) = j_{\gamma-1/2}\big(2\sqrt{\chi_0}\big) + O(\tau),$$

and

$$(1-\phi)^{2\gamma}F\Big(\frac{1}{2}+\gamma, \frac{1}{2}+\gamma; 1;\ \phi\Big) = F\Big(\frac{1}{2}-\gamma, \frac{1}{2}-\gamma; 1;\ \phi\Big)$$

$$= \frac{\Gamma(2\gamma)}{\Gamma^2(\gamma+1/2)} + \begin{cases} O\big((1-\phi)^{2\gamma}\big), & 0<\gamma<\frac{1}{2}, \\ O(1-\phi), & \gamma>\frac{1}{2} \end{cases}$$

for $\tau \to 0+$, the assertion of Lemma 3(i) then follows.

Concerning part (ii), term by term differentiation for $\phi<1$ is justified by Lemma 1, (3.3), so that

(5.7)

$$\frac{\partial}{\partial\tau}a_L^{\gamma}(\tau)=\left(\gamma-\frac{1}{2}\right)\psi^{\gamma-3/2}\psi'(\tau)\sum_{n=0}^{\infty}S\left(n,\frac{1}{2}-\gamma;\ \phi,\chi\right)\frac{(1/2-\gamma)_n(1/2-\gamma)_n}{n!n!}\phi^n$$
$$+\psi^{\gamma-1/2}\phi'(\tau)\sum_{n=1}^{\infty}S\left(n,\frac{1}{2}-\gamma;\ \phi,\chi\right)\frac{(1/2-\gamma)_n(1/2-\gamma)_n}{n!n!}\phi^{n-1}$$
$$+\psi^{\gamma-1/2}\sum_{n=0}^{\infty}\frac{\partial}{\partial\tau}S\left(n,\frac{1}{2}-\gamma;\ \phi,\chi\right)\frac{(1/2-\gamma)_n(1/2-\gamma)_n}{n!n!}\phi^n$$

and

$$\frac{\partial}{\partial\tau}S\left(n,\frac{1}{2}-\gamma;\ \phi,\chi\right)$$
$$=\phi'(\tau)\left(n+\frac{1}{2}-\gamma\right)\sum_{k=0}^{\infty}kF\left(n+\frac{3}{2}-\gamma,k+1;\ n+k+2;\ \phi\right)\frac{n!(-\chi)^k}{k!(k+n+1)!}$$
$$-\chi'(\tau)\sum_{k=0}^{\infty}F\left(n+\frac{1}{2}-\gamma,k+1;\ n+k+2;\ \phi\right)\frac{n!(-\chi)^k}{k!(k+n+1)!}.$$

An index transformation in the second term of (5.7), together with the relation

$$\left(n+\frac{1}{2}-\gamma\right)F\left(n+\frac{3}{2}-\gamma,k;\ n+k+2;\ \phi\right)+kF\left(n+\frac{3}{2}-\gamma,k+1;\ n+k+2;\ \phi\right)$$
$$=(n+k+1)F\left(n+\frac{3}{2}-\gamma,k;\ n+k+1;\ \phi\right)-\left(\gamma+\frac{1}{2}\right)F\left(n+\frac{3}{2}-\gamma,k;\ n+k+2;\ \phi\right),$$

then give

$$\frac{\partial}{\partial\tau}a_L^{\gamma}(\tau)$$
$$=\left(\gamma-\frac{1}{2}\right)\psi^{\gamma-3/2}\psi'(\tau)\sum_{n=0}^{\infty}S\left(n,\frac{1}{2}-\gamma;\ \phi,\chi\right)\frac{(1/2-\gamma)_n(1/2-\gamma)_n}{n!n!}\phi^n$$
$$+\left(\frac{1}{2}-\gamma\right)\psi^{\gamma-1/2}\phi'(\tau)\sum_{n=0}^{\infty}S\left(n,\frac{3}{2}-\gamma;\ \phi,\chi\right)\frac{(1/2-\gamma)_n(3/2-\gamma)_n}{n!n!}\phi^n$$
$$+\left(\gamma-\frac{1}{2}\right)\left(\gamma+\frac{1}{2}\right)\psi^{\gamma-1/2}\phi'(\tau)\sum_{n=0}^{\infty}S\left(n+1,\frac{1}{2}-\gamma;\ \phi,\chi\right)\frac{(1/2-\gamma)_n(3/2-\gamma)_n}{(n+1)!n!}\phi^n$$
$$-\psi^{\gamma-1/2}\chi'(\tau)\sum_{n=0}^{\infty}\left\{\sum_{k=0}^{\infty}F\left(n+\frac{1}{2}-\gamma,k+1;\ n+k+2;\ \phi\right)\frac{n(-\chi)^k}{k!(k+n+1)!}\right\}$$
$$\cdot\frac{(1/2-\gamma)_n(1/2-\gamma)_n}{n!n!}\phi^n$$
$$=A_1+A_2+A_3+A_4,$$

say. When estimating the four terms separately, they turn out to be of the same order if $\gamma > \frac{1}{2}$, while for $0 \leqq \gamma < \frac{1}{2}$ the second term turns out to be the main one. Indeed, since

$$|A_1| \leqq \psi^{\gamma-3/2}\left|\left(\gamma-\frac{1}{2}\right)\psi'(\tau)\left|F\left(\frac{1}{2}-\gamma, \frac{1}{2}-\gamma; 1; \phi\right)\right.\right.,$$

$$|A_3| \leqq C_\gamma \psi^{\gamma-1/2}\left|\left(\gamma-\frac{1}{2}\right)\left(\gamma+\frac{1}{2}\right)\phi'(\tau)F\left(\frac{1}{2}-\gamma, \frac{3}{2}-\gamma; 2; \phi\right)\right|$$

where the constant C_γ equals 1 for $0 \leqq \gamma < \frac{1}{2}$, and

$$|A_4| \leqq \psi^{\gamma-1/2}|\chi'(\tau)| \sum_{n=0}^{\infty} \int_0^1 |j_0(2\sqrt{\chi t})|(1-t)^n(1-t\phi)^{-n-1/2+\gamma}dt \cdot \frac{(1/2-\gamma)_n(1/2-\gamma)_n}{n!n!}\phi^n$$

$$\leqq \psi^{\gamma-1/2}|\chi'(\tau)|F\left(\frac{1}{2}-\gamma, \frac{1}{2}-\gamma; 1; \phi\right) \cdot \begin{cases} \dfrac{1}{\gamma+1/2}, & 0 \leqq \gamma < \dfrac{1}{2}, \\ 1, & \gamma \geqq \dfrac{1}{2} \end{cases}$$

it follows for $i = 1, 3, 4$ that

$$A_i = \begin{cases} O\left(\log \dfrac{1}{\tau}\right), & \gamma = 0, \\ O(1), & \gamma > 0 \end{cases} \qquad (\tau \to 0+). \tag{5.8}$$

For $\gamma > \frac{1}{2}$, A_2 is of the same order since, in view of (3.3),

$$|A_2| \leqq \left(\gamma-\frac{1}{2}\right)\psi^{\gamma-1/2}|\phi'(\tau)| \cdot \sum_{n=0}^{\infty} \frac{|(1/2-\gamma)_n(3/2-\gamma)_n|}{n!n!}\phi^n \cdot \begin{cases} 1+\dfrac{\chi}{\gamma^2-1/4}, & \dfrac{1}{2} < \gamma < 1, \\ 1, & \gamma \geqq 1 \end{cases} \tag{5.9}$$

$$= O(1) \qquad (\tau \to 0+).$$

It remains to consider A_2 in case $0 \leqq \gamma < \frac{1}{2}$. Euler's transformation and integral representation give

$$S\left(n, \frac{3}{2}-\gamma; \phi, \chi\right)$$

$$= (1-\phi)^{\gamma-1/2} \sum_{k=0}^{\infty} F\left(k+\gamma-\frac{1}{2}, n+1; k+n+1; \phi\right)\frac{n!(-\chi)^k}{k!(k+n)!}$$

$$= 1 + (1-\phi)^{\gamma-1/2}\frac{n!}{(3/2-\gamma)_n\Gamma(3/2-\gamma)} \cdot \sum_{k=1}^{\infty} \int_0^1 t^{k+\gamma-3/2}(1-t)^{n-\gamma+1/2}(1-t\phi)^{-n-1}dt\frac{(-\chi)^k}{\Gamma(k+\gamma-1/2)k!}.$$

Inserting this into A_2 and interchanging the order of summation we have

$$A_2=\left(\frac{1}{2}-\gamma\right)\psi^{\gamma-1/2}\phi'(\tau)$$
$$\cdot\left[F\left(\frac{1}{2}-\gamma,\frac{3}{2}-\gamma;1;\phi\right)+\frac{1}{\Gamma(3/2-\gamma)}(1-\phi)^{\gamma-1/2}\right.$$
$$\cdot\sum_{k=1}^{\infty}\left\{\sum_{n=0}^{\infty}\frac{(1/2-\gamma)_n}{n!}\int_0^1 t^{k+\gamma-3/2}(1-t)^{n-\gamma+1/2}(1-t\phi)^{-n-1}dt\phi^n\right\}$$
$$\left.\cdot\frac{(-\chi)^k}{\Gamma(k+\gamma-1/2)k!}\right],$$

and, by (5.6), the sum in curly brackets equals

$$(1-\phi)^{\gamma-1/2}\int_0^1 t^{k+\gamma-3/2}(1-t)^{-\gamma+1/2}(1-t\phi)^{-\gamma-1/2}dt$$
$$=(1-\phi)^{\gamma-1/2}\frac{\Gamma(k+\gamma-1/2)\Gamma(3/2-\gamma)}{\Gamma(k+1)}F\left(\gamma+\frac{1}{2},k+\gamma-\frac{1}{2};k+1;\phi\right).$$

Using the transformation

$$F\left(\frac{1}{2}-\gamma,\frac{3}{2}-\gamma;1;\phi\right)=(1-\phi)^{2\gamma-1}F\left(\gamma+\frac{1}{2},\gamma-\frac{1}{2};1;\phi\right)$$

and applying Euler's integral representation once more, one obtains

$$A_2=\left(\frac{1}{2}-\gamma\right)\psi^{\gamma-1/2}\phi'(\tau)(1-\phi)^{2\gamma-1}\sum_{k=0}^{\infty}F\left(\gamma+\frac{1}{2},k+\gamma-\frac{1}{2};k+1;\phi\right)\frac{(-\chi)^k}{\Gamma^2(k+1)}.$$

Using now (3.2a), $1-\phi=\psi^{-1}\xi\tau xt$ and $\phi'(\tau)=-\psi_0^{-1}\xi xt+O(\tau)$, $\tau\to 0+$, we arrive at

$$A_2=\frac{\Gamma(1-2\gamma)}{\Gamma^2(1/2-\gamma)}\psi^{\gamma-1/2}\phi'(\tau)(1-\phi)^{2\gamma-1}j_{-\gamma-1/2}(2\sqrt{\chi})+\begin{cases}O\left(\log\frac{1}{\tau}\right), & \gamma=0,\\ O(1), & 0<\gamma<\frac{1}{2},\end{cases}$$

(5.10)

$$=-\frac{\Gamma(1-2\gamma)}{\Gamma^2(1/2-\gamma)}\psi_0^{-\gamma-1/2}(\xi xt)^{2\gamma}j_{-\gamma-1/2}\left(2\sqrt{\chi_0}\right)\tau^{2\gamma-1}+\begin{cases}O\left(\log\frac{1}{\tau}\right), & \gamma=0,\\ O(1), & 0<\gamma<\frac{1}{2}\end{cases}$$

as $\tau\to 0+$. By (5.8), (5.9), and (5.10), the proof of Lemma 3 is complete. □

Applying Lemma 3 for $\gamma=|\alpha|$ to the right-hand side of (5.4) and observing (5.5), the desired asymptotic expansions now follow:

$$G_L^\alpha(\tau)=-\frac{\Gamma(2\alpha+1)}{\Gamma^2(\alpha+1/2)}\xi(xt)^{-2\alpha}\psi_0^{\alpha-1/2}j_{\alpha-1/2}\left(2\sqrt{\chi_0}\right) \tag{5.11}$$

$$+\begin{cases} O(\tau), & \alpha\geqq\frac{1}{2},\\ O(\tau^{2\alpha}), & 0<\alpha<\frac{1}{2},\\ O\left(\tau\log\frac{1}{\tau}\right), & \alpha=0,\\ O(\tau^{2\alpha+1}), & -\frac{1}{2}<\alpha<0 \end{cases} \qquad (\tau\to 0+),$$

$$G_L^{-1/2}(\tau)=\frac{1}{2}\xi xt\,j_1\left(2\sqrt{\chi_0}\right)+O(\tau) \qquad (\tau\to 0+),$$

where $\psi_0=\chi_0=\frac{1}{16}[(x+t)^2-\xi^2][\xi^2-(x-t)^2]$. In view of (5.1) we thus have found that

(5.12)

$$w_L^\alpha(x,t,\xi)=\begin{cases} \dfrac{\Gamma(2\alpha+1)}{\Gamma^2(\alpha+1/2)}\xi(2xt)^{-2\alpha}(\rho(x,t,\xi))^{2\alpha-1}j_{\alpha-1/2}(\rho(x,t,\xi)), & \alpha>-\frac{1}{2},\\ -\dfrac{1}{4}\xi xt\,j_1(\rho(x,t,\xi)), & \alpha=-\frac{1}{2}, \end{cases}$$

where ρ is given by (1.7).

Proof of Theorem 1. Using (2.4) and Legendre's duplication formula, the kernel representation (1.6) is an immediate consequence of (5.12). □

6. A comparison between the Laguerre and the Hankel translations. There is a deep relation between the geometrical interpretations of both translations. In fact the Laguerre translation plays the same role in the rotation invariant case of the Weyl transform on a higher dimensional Heisenberg group as the Hankel translation does with respect to the rotation invariant case of the ordinary Fourier transform in several variables [24].

But while the Hankel translation is a positive operator, the Laguerre translation obviously is not. Nevertheless, it follows from the estimate (3.5) of the Bessel functions that, for each $\alpha\geqq 0$, the respective translation kernels are related by

$$|K_L^\alpha(x,t,z)|\leqq K_H^\alpha(x,t,z) \qquad (0<x,t,z<\infty).$$

Consequently, the operator norms of the translations satisfy

$$\|T_{L,t}^\alpha\|_{[L^1_{w(2\alpha+1)}]}\leqq\|T_{H,t}^\alpha\|_{[L^1_{w(2\alpha+1)}]}=1 \qquad (\alpha\geqq 0,\ t\geqq 0),$$

and, in fact, it can be shown that equality holds [12]. To our opinion this is a surprising result which seems to indicate that even nonpositive translation operators may show a favourable norm behavior.

In addition to the numerous relations between the Laguerre and Hankel cases mentioned in the preceding paragraphs, it is to be noted that the Riemann functions themselves are closely related to each other by

$$|A_L^\alpha(\xi,\tau;\,x,t)|\leqq A_H^\alpha(\xi,\tau;\,x,t) \tag{6.1}$$

provided $\alpha \geqq -\frac{1}{2}$ and $(\xi,\tau)\in\overline{\Delta_{xt}}$. This follows from (4.9), (4.13) by observing that the factors $S(n,\frac{1}{2}-|\alpha|;\phi,\chi)$ are uniformly bounded by 1 in view of Lemma 1.

We finally mention two different formal representations of the translation kernels as a "triple product integral" and a "triple product sum", which reflect the formal difference arising from the continuous and discrete spectrum of the S.-L. DE (1.4) for $q\equiv 0$ and $q(x)=x^2$, respectively. These are

$$K_H^\alpha(x,t,z)=[2^\alpha\Gamma(\alpha+1)]^{-2}\int_0^\infty j_\alpha(xu)\,j_\alpha(tu)\,j_\alpha(zu)\,u^{2\alpha+1}\,du,$$

$$K_L^\alpha(x,t,z)=\sum_{k=0}^{\infty}\frac{2\Gamma(k+\alpha+1)}{\Gamma^2(\alpha+1)\Gamma(k+1)}\,y_k^\alpha(x)\,y_k^\alpha(t)\,y_k^\alpha(z)$$

if $|x-t|<z<x+t$. This can also be compared with the kernel representation of the Jacobi translation given by Gasper [11, (2.2)] as a sum over triple products of Jacobi polynomials as well as an integral of a triple product of Bessel functions.

Acknowledgment. The author would like to thank Professor E. Görlich for helpful comments and suggestions.

REFERENCES

[1] R. ASKEY, *Orthogonal polynomials and positivity*, in Studies in Applied Mathematics, Wave Propagation and Special Functions, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1970, pp. 64–85.

[2] S. BOCHNER, *Sturm–Liouville and heat equations whose eigenfunctions are ultraspherical polynomials or associated Bessel functions*, Proc. Conference on Differential Equations, Maryland 1955, J. B. Diaz and L. E. Payne, eds., Maryland Univ. Press, MD, 1956, pp. 23–48.

[3] B. L. J. BRAAKSMA, *A singular Cauchy problem and generalized translations*, International Conference on Differential Equations, Los Angeles 1974, H. A. Antosiewicz, ed., Academic Press, New York, 1975, pp. 40–52.

[4] B. L. J. BRAAKSMA AND H. S. V. DE SNOO, *Generalized translation operators associated with a singular differential operator*, Proc. Conference on Theory of Ordinary and Partial Differential Equations, Dundee 1974, B. D. Sleeman and I. M. Michaels, eds., Lecture Notes in Mathematics 415, Springer, Berlin, 1974, pp. 62–77.

[5] H. CHÉBLI, *Sur la positivité des opérateurs de "translation généralisée" associés à un operateur de Sturm–Liouville sur* $]0,\infty[$, C. R. Acad. Sci. Paris, 275 (1972), pp. 601–604.

[6] F. M. CHOLEWINSKI AND D. T. HAIMO, *Classical analysis and the generalized heat equation*, SIAM Rev., 10 (1968), pp. 67–80.

[7] J. DELSARTE, *Sur une extension de la formule de Taylor*, J. Math. Pures Appl., 17 (1936), pp. 213–231.

[8] _____, *Une extension nouvelle de la théorie de fonctions presque périodique de Bohr*, Acta Math., 69 (1938), pp. 259–317.

[9] A. ERDÉLYI, et al., *Higher Transcendental Functions*, Vol. 1, McGraw-Hill, New York, 1953.

[10] P. R. GARABEDIAN, *Partial Differential Equations*, John Wiley, New York, 1964.

[11] G. GASPER, *Positivity and the convolution structure for Jacobi series*, Ann. of Math., 93 (1971), pp. 112–118.

[12] E. GÖRLICH AND C. MARKETT, *A convolution structure for Laguerre series*, Indag. Math., 44 (1982), pp. 161–171.

[13] D. L. GUY, *Hankel multiplier transformations and weighted p-norms*, Trans. Amer. Math. Soc., 95 (1960), pp. 137–189.

[14] P. HENRICI, *A survey of I. N. Vekua's theory of elliptic differential equations with analytic coefficients*, Z. Angew. Math. Phys., 8 (1957), pp. 169–203.

[15] I. I. HIRSCHMAN, JR., *Variation diminishing Hankel transforms*, J. Analyse Math., 8 (1960/61), pp. 307–336.

[16] T. H. KOORNWINDER, *The addition formula for Jacobi polynomials*, III. *Completion of the proof*, Math. Centrum Amsterdam Rep., TW 135, 1972.

[17] _____, *Jacobi polynomials and their two-variable analogues*, Thesis, University of Amsterdam, 1974.
[18] _____, *Jacobi polynomials*, II. *An analytic proof of the product formula*, this Journal, 5 (1974), pp. 125–137.
[19] _____, *Jacobi polynomials*, III. *An analytic proof of the addition formula*, this Journal, 6 (1975), pp. 533–543.
[20] _____, *The addition formula for Laguerre polynomials*, this Journal, 8 (1977), pp. 535–540.
[21] E. LANCKAU, *Zur Lösung gewisser partieller Differentialgleichungen mittels parameterabhängiger Bergmann-Operatoren*, Nova Acta Leopoldina Nr. 201, 36 (1971), 46 pp.
[22] B. M. LEVITAN, *Generalized Translation Operators and Some of Their Applications*, Israel Program for Scientific Translations, Jerusalem, 1964.
[23] C. MARKETT, *Mean Cesàro summability of Laguerre expansions and norm estimates with shifted parameter*, Anal. Math., 8 (1982), pp. 19–37.
[24] J. PEETRE, *The Weyl transform and Laguerre polynomials*, Le Matematiche, 27 (1972), pp. 301–323.
[25] A. POVZNER, *On differential equations of Sturm–Liouville type on a half axis*, Amer. Math. Soc. Transl., 5 (1950), pp. 24–101; Mat. Sbornik (N.S.), 23(65) (1948), pp. 3–52.
[26] J. PÜNGEL, *Riemannfunktionen mit speziellen Potenzreihenentwicklungen*, Appl. Anal., 11 (1981), pp. 199–210.
[27] B. RIEMANN, *Über die Fortpflanzung ebener Luftwellen von endlicher Schwingungsweite*, Gesammelte Math. Werke, Leipzig, 1892, pp. 156–181.
[28] R. L. ŠAPIRO, *Special functions related to representations of the group $SU(n)$, of class I with respect to $SU(n-1)$ ($n \geqq 3$)*, Izv. Vyss. Učebn. Zaved. Matematika, 71 (1968), pp. 97–107. (In Russian.)
[29] L. J. SLATER, *Generalized Hypergeometric Functions*, Cambridge Univ. Press, Cambridge, 1966.
[30] G. SZEGÖ, *Orthogonal Polynomials*, 4th ed., AMS Colloquium Publications, vol. 23, American Mathematical Society, Providence, RI, 1975.
[31] H. WALLNER, *Riemann-Funktionen mit mehreren Hilfsveränderlichen*, Berichte Math.-Stat. Sektion Forschungszentrum Graz, Nr. 174, Graz 1981, 27 pp.
[32] G. N. WATSON, *Another note on Laguerre polynomials*, J. London Math. Soc., 14 (1939), pp. 19–22.
[33] _____, *A Treatise on the Theory of Bessel Functions*, 2nd ed., Cambridge Univ. Press, Cambridge, 1962.

SIAM J. MATH. ANAL.
Vol. 17, No. 4, July 1986

019

AN ANALYTICAL EXPRESSION FOR COEFFICIENTS ARISING WHEN IMPLEMENTING A TECHNIQUE FOR INDEFINITE INTEGRATION OF PRODUCTS OF SPECIAL FUNCTIONS*

JEAN C. PIQUETTE†

Abstract. In an earlier article [*Technique for evaluating indefinite integrals involving products of certain special functions*, Jean C. Piquette and A. L. Van Buren, SIAM J. Math. Anal., 15 (1984), pp. 845–855], a new analytical technique for evaluating indefinite integrals involving special functions was described. 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. The present article gives an analytical expression for the functional coefficients arising in the coupled set. Thus, the process of obtaining the coupled set is reduced to evaluation of the analytical expression. This formula is therefore valuable in creating a general algorithm for analytically evaluating indefinite integrals of special functions. Such an algorithm could be used to mechanically generate the coupled differential equations, and could be readily implemented on a digital computer using a symbolic manipulation program.

Key words. antidifferentiation, ordinary differential equations, systems series expansion

AMS(MOS) subject classifications. Primary 34A30, 41A58

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 the ith type of special function 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 $\prod R_\mu^{(i)}$ are both assumed continuous (or with at most a finite number of discontinuities) over an interval $[x_1,x_2]$, insuring that the integral I exists in the same interval. The technique is a generalization of one used by Sonine [2] (described by Watson [3]) to evaluate indefinite integrals involving products of Bessel functions. Reference [1] extended the technique to include most of the special functions of physics, including Legendre functions, Hermite functions, Laguerre funtions, etc. The technique replaces the integral which 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.

*Received by the editors January 22, 1985, and in revised form June 30, 1985.
† Naval Research Laboratory, Orlando, Florida 32856-8337.

The method of [1] assumes that the integral I of (1) may be represented by the expression

$$I=\sum_{p_1=0}^{1}\sum_{p_2=0}^{1}\cdots\sum_{p_m=0}^{1}A_{p_1p_2,\cdots,p_m}(x)\prod_{i=1}^{m}R^{(i)}_{\mu_i+p_i}(x), \tag{3}$$

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 notations $A_p\equiv A_{p_1,p_2,\cdots,p_m}(x)$ and $B_{pq}\equiv B_{p_1,p_2,\cdots,p_m,q_1,q_2,\cdots,q_m}(x)$ have been used. Also, the notation $\Sigma_{\{q\}}$ 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 described in [1] as being known functions resulting from repeated applications of the relations of (2) and the regrouping of terms in the form $\Pi_{i=1}^{m}R^{(i)}_{\mu_i+p_i}$. As of the writing of [1], no general analytical expression for the coefficients B_{pq} was available. However, due to the advent of symbolic manipulation computer programs (such as MACSYMA, SMP, REDUCE, MAPLE, etc.), such a formula would reduce the problem of generating the coupled set (4) from the indefinite integral of (1) to a mechanical process (i.e., evaluation of a formula). Since a particular solution to the coupled set may be easier to obtain than directly performing the indefinite integral, the availability of a formula to generate the required coupled set in an algorithmic way is of interest. This formula is obtained in the present article.

2. An analytical expression for the coefficients B_{pq}. Rather than working directly with the recurrence relations (2), it is more convenient to use the equivalent set

$$DR^{(i)}_{\mu}(x)=a'_{\mu}(x)R^{(i)}_{\mu}(x)+b'_{\mu}(x)R^{(i)}_{\mu+1}(x), \tag{5a}$$

$$DR^{(i)}_{\mu+1}(x)=c'_{\mu}(x)R^{(i)}_{\mu}(x)+d'_{\mu}(x)R^{(i)}_{\mu+1}(x), \tag{5b}$$

where

$$a'_{\mu}(x)\equiv c_{\mu}(x)-\frac{a_{\mu}(x)d_{\mu}(x)}{b_{\mu}(x)}, \qquad c'_{\mu}(x)\equiv d_{\mu+1}(x),$$

$$b'_{\mu}(x)\equiv d_{\mu}(x)/b_{\mu}(x), \qquad d'_{\mu}(x)\equiv c_{\mu+1}(x).$$

The special case where $b_{\mu}=0$ (discussed extensively in [1]) is excluded here.

The desired expression for B_{pq} is now obtained in the following way: The logarithmic derivative of the expression $\Pi_{i=1}^{m}R^{(i)}_{\mu_i+p_i}(x)$ is first computed and simplified using recurrence relations (5). It is next necessary to express the quantity

$$\sum_{\{p\}}A_pD\prod_{i=1}^{m}R^{(i)}_{\mu_i+p_i}(x)$$

in terms of expressions of the form $\Pi_{i=1}^{m} R^{(i)}_{\mu_i+p_i}(x)$. The desired expression for B_{pq} is obtained by comparing the result of the above computation with second term on the right-hand side of [1, eq. (10)]. The result is

$$(6)\qquad B_{pq}=\sum_{j=1}^{m}\left[\left(a'_{\mu_j}(x)\prod_{l=1}^{m}\delta_{q_l,p_l}+c'_{\mu_j}(x)\prod_{l=1}^{m}\delta_{q_l,p_l+\delta_{lj}}\right)\delta_{p_j,0}\right.$$
$$\left.+\left(b'_{\mu_j}(x)\prod_{l=1}^{m}\delta_{q_l,p_l-\delta_{lj}}+d'_{\mu_j}(x)\prod_{l=1}^{m}\delta_{q_l,p_l}\right)\delta_{p_j,1}\right],$$

where a'_μ, b'_μ, c'_μ, and d'_μ are defined above. Although the number of coefficients B_{pq} becomes large even for small values of m (this number being equal to 2^{2m}), this presents no difficulty to the new symbol manipulation computer programs which routinely evaluate symbolic expressions containing thousands of terms. [1] discusses the procedure necessary to obtain a series solution (involving a *single summation index*) for the unknown integral. This process could also be implemented mechanically using a symbol manipulation program. In fact, I have recently implemented [1, eq. (10)], replacing B_{pq} with (6) above, using the SMP© computer program. This implementation has enabled generation of the relevant differential equations to occur in a few minutes, compared with the several hours normally required when implemented using paper and pencil. Hence, combining the technique of [1] and the formula of (6) above, it should generally be possible to obtain a simple series representation for integrals of the type (1). Of course, if a closed-form particular solution of the coupled set (4) is found, then a closed-form analytical expression for the integral of (1) results via (3).

REFERENCES

[1] J. C. PIQUETTE AND A. L. VAN BUREN, *Technique for evaluating indefinite integrals involving products of certain special functions*, this Journal, 15 (1984), pp. 845–855.

[2] N. J. SONINE, Math. Ann., XVI (1880), pp. 1–80.

[3] G. N. WATSON, *A Treatise on the Theory of Bessel Functions*, Cambridge Univ. Press, London, 1966, pp. 132–134.

SIAM J. MATH. ANAL.
Vol. 17, No. 5, September 1986

RECURRENCE RELATIONS FOR THE COEFFICIENTS IN JACOBI SERIES SOLUTIONS OF LINEAR DIFFERENTIAL EQUATIONS*

STANISŁAW LEWANOWICZ†

Abstract. A method is presented for obtaining recurrence relations for the coefficients in Jacobi series solutions of linear ordinary differential equations with polynomial coefficients.

Key words. Jacobi series, Jacobi coefficients, recurrence relations, difference operators, linear differential equation

AMS(MOS) subject classifications. Primary 42C10, 39A70, 65L05, 65L10

1. Introduction. A function defined in the interval $\langle -1,1\rangle$ and satisfying the required conditions (see, e.g., [3, v. 1, §10.19] or [11, v. 1, §8.3]) may be expanded into a series uniformly convergent in this interval with respect to the Jacobi polynomials (in short: the *Jacobi series*):

$$f=\sum_{k=0}^{\infty} a_k^{(\alpha,\beta)}[f]P_k^{(\alpha,\beta)} \qquad (\alpha>-1,\beta>-1). \tag{1.1}$$

Here $P_k^{(\alpha,\beta)}$ is the usual notation for the kth Jacobi polynomial, and

$$a_k^{(\alpha,\beta)}[f]=\frac{(2k+\lambda)k!\Gamma(k+\lambda)}{2^{\lambda}\Gamma(k+\alpha+1)\Gamma(k+\beta+1)}\int_{-1}^{1}(1-x)^{\alpha}(1+x)^{\beta}P_k^{(\alpha,\beta)}(x)f(x)\,dx \tag{1.2}$$

$$(k=0,1,\cdots)$$

where

$$\lambda=\alpha+\beta+1.$$

Some alternative forms for the coefficients (1.2) are available for many elementary and special functions (see, e.g., [11, v. 2, §9.3]). In particular, if f belongs to the hypergeometric family, $a_k^{(\alpha,\beta)}[f]$ is also of this type and obeys a difference equation of the form

$$\sum_{j=0}^{r} A_j(k)\psi_{k+j}=B(k), \tag{1.3}$$

in which A and B are rational in k [11, v. 2, §12.4], [10]. In this case, equation (1.3) serves as the basis for a very efficient numerical procedure for the calculation of $a_k^{(\alpha,\beta)}[f]$ (see [11, v. 2, §12.5] or [17, Ex. 7.2]).

A simple and universal method for deriving a recurrence relation for the coefficients $a_k^{(\alpha,\beta)}[f]$ may be applied in the case where the function f satisfies the linear differential equation

$$\sum_{m=0}^{n} p_m f^{(m)}=q \tag{1.4}$$

* Received by the editors April 15, 1985.
† Institute of Computer Science, University of Wrocław, 51-151 Wrocław, Poland.

of order n, with suitable initial or boundary conditions. In (1.4) p_0, $p_1, \cdots, p_n$ are polynomials, and the coefficients $a_k^{(\alpha,\beta)}[q]$ are known. The main idea of this method should be ascribed to Clenshaw [1]. We restate Clenshaw's result briefly. We assume the *Chebyshev series* expansion for f:

$$f=\frac{1}{2}t_0[f]T_0+\sum_{k=0}^{\infty} t_k[f]T_k \tag{1.5}$$

and similar expansions for the derivatives, where T_k is the Chebyshev polynomial

$$T_k(x)=\frac{k!}{(1/2)_k}P_k^{(-1/2,-1/2)}(x). \tag{1.6}$$

Here we use the Pochhammer symbol

$$(a)_k=\Gamma(a+k)/\Gamma(a).$$

The following identities can be deduced from basic difference properties of the Chebyshev polynomials [1]:

$$t_{k-1}[f^{(m+1)}]-t_{k+1}[f^{(m+1)}]=2kt_k[f^{(m)}], \tag{1.7}$$

$$t_k[x^s f^{(m)}]=2^{-s}\sum_{j=0}^{s}\binom{s}{j}t_{k-s+2j}[f^{(m)}], \tag{1.8}$$

for nonnegative integers m and s. The above equations are then applied to the differential equation (1.4) to obtain a system of finite difference equations for the coefficients $\{t_k[f^{(m)}]\}$ $(m=0,1,\cdots,n)$. Many authors (Fox [4], Fox and Parker [5], Geddes [6], Horner [7], Morris and Horner [13], Olaofe [14], Paszkowski [15]) have proposed modifications and improvements of Clenshaw's method, which lead to a single difference equation or recurrence relation for the principal coefficients $\{t_k[f]\}$.

Paszkowski ([15, §13]) has raised the problem of constructing the recurrence relation which has the lowest order among all such relations following from (1.4), (1.7) and (1.8). The complete solution to this problem, even in the more general case of the *Gegenbauer series* expansion

$$f=\sum_{k=0}^{\infty} g_k^{(\nu)}[f]C_k^{(\nu)} \qquad (\nu>-1/2), \tag{1.9}$$

where $C_k^{(\nu)}$ is the Gegenbauer polynomial

$$\begin{aligned} C_k^{(\nu)}&=\frac{(2\nu)_k}{(\nu+1/2)_k}P_k^{(\nu-1/2,\nu-1/2)} \qquad (\nu\neq 0),\\ C_k^{(0)}&=\lim_{\nu\to 0}\nu^{-1}C_k^{(\nu)}, \end{aligned} \tag{1.10}$$

was given by Lewanowicz [8]. A generalization of Clenshaw's method, in which (1.6) was replaced by (1.9), has been given by Elliott [2].

The case of the unsymmetric Jacobi polynomial expansion (1.1), for which $\alpha\neq\beta$, was discussed in [9]; the proposed optimum algorithm provides a recurrence relation of minimum order among all such equations which can be obtained from the differential equation (1.4), using basic difference and differential properties of the Jacobi polynomials (see (3.1), (3.2)).

In the present paper we describe another method for constructing a recurrence relation for the coefficients (1.2). The new method, though not optimum in general, seems to have some important advantages over the algorithms of papers [8] and [9]. First, the recurrence relation and its order are given by explicit formulae. Second, the components of these formulae are expressed in terms of the coefficients of the differential equation

$$\sum_{m=0}^{n} (q_m f)^{(m)} = q, \tag{1.11}$$

where

$$q_m = \sum_{j=m}^{n} (-1)^{j-m} \binom{j}{m} p_j^{(j-m)}, \tag{1.12}$$

which is equivalent to (1.4) (see, e.g., [15], p. 231). Last, the new procedure involves much less computational effort. Also, it could well be programmed in a language for symbolic computation; in this connection, see [6] and [16].

The main result of the paper is given in §4. The special cases of Gegenbauer and Chebyshev expansions are discussed in §5. Section 6 contains an illustrative example.

In the sequel we shall use the notation [9]

$$b_k[f] \equiv b_k^{(\alpha,\beta)}[f] = \frac{\Gamma(k+\alpha+1)}{\Gamma(k+\lambda)(2k+\lambda-1)_3} a_k^{(\alpha,\beta)}[f]. \tag{1.13}$$

We call $b_k[f]$ the *Jacobi coefficients* of the function f. It will be convenient to use coefficients with negative indices. We assume that if $\alpha \neq \beta$, or $\alpha = \beta$ but $2\alpha+1$ is not an integral $\geqq 0$, then

$$b_{-k}^{(\alpha,\beta)}[f] = 0 \quad \text{for } k = 1,2,\cdots, \tag{1.14}$$

and if $\alpha = \beta$ and $2\alpha+1 = m$ is a nonnegative integer, we define [2]

$$b_{-k}^{(\alpha,\alpha)}[f] = \begin{cases} 0 & \text{for } k = 1,2,\cdots,m-1, \\ b_{k-m}^{(\alpha,\alpha)}[f] & \text{for } k \geqq m. \end{cases} \tag{1.15}$$

The quantities

$$c_k[f] \equiv c_k^{(\nu)}[f] = 2\left(k+\nu-\frac{1}{2}\right) b_k^{(\nu-1/2,\nu-1/2)}[f] \tag{1.16}$$

are called the *Gegenbauer coefficients* of f. It can be seen that

$$c_k^{(\nu)}[f] = \begin{cases} \dfrac{\sqrt{\pi}}{\Gamma(\nu)(k+\nu)} g_k^{(\nu)}[f] & (\nu \neq 0), \\ \dfrac{\sqrt{\pi}}{4} t_k[f] & (\nu = 0). \end{cases}$$

2. Difference operators. The results given in the following sections are expressed in terms of a certain type of linear operator. Let $\mathscr{S}$ denote the linear space of all "doubly infinite" sequences of complex numbers, with addition of sequences and scalar multiplication defined as usual. Obviously $\mathscr{S}$ is the space of all complex-valued func-

tions defined on the set of all integers. Let $\mathscr{S}_{\text{rat}}$ denote the set of all *rational* functions $s\in\mathscr{S}$.

Consider the set $\mathscr{S}^*$ of all linear operators mapping $\mathscr{S}$ into itself. For $T\in\mathscr{S}^*$ and $\{z_k\}\in\mathscr{S}$, we denote the kth coordinate of the sequence $T\{z_k\}\in\mathscr{S}$ by Tz_k, so that $T\{z_k\}=\{Tz_k\}$. The *zero operator*, the *identity operator* and the *mth shift operator* in $\mathscr{S}^*$ are denoted by θ, I and E^m, respectively. Then we have

$$Iz_k=z_k,\quad \theta z_k=0,\quad E^m z_k=z_{k+m} \tag{2.1}$$

for every $\{z_k\}\in\mathscr{S}$. Clearly, $E^0=I$.

Let $\mathscr{L}$ be the set of all operators $L\in\mathscr{S}^*$ such that

$$L=\sum_{j=0}^{r}\lambda_j(k)E^{u+j}, \tag{2.2}$$

where $r\geqq 0$ and u are integers, and $\lambda_0,\lambda_1,\cdots,\lambda_r\in\mathscr{S}_{\text{rat}}$. Every nonzero operator $L\in\mathscr{L}$ can be expressed in the form (2.2) with $\lambda_0\not\equiv 0$ and $\lambda_r\not\equiv 0$. The number $r=r(L)$ is referred to as the *order* of the operator L, while λ_j are called the coefficients of L. The elements of the set $\mathscr{L}$ are known as *difference operators*.

Let $L\in\mathscr{L}$ be defined by (2.2) and let $M\in\mathscr{L}$ be such that

$$M=\sum_{j=0}^{t}\mu_j(k)E^{v+j}.$$

We define the *product* of L and M to be the operator

$$LM=\sum_{i=0}^{r}\lambda_i(k)\sum_{j=0}^{t}\mu_j(k+u+i)E^{u+v+i+j}.$$

It can be seen that under this definition of multiplication, with addition defined in a natural manner, $\mathscr{L}$ forms a ring with identity I.

Let $L\in\mathscr{L}$ and $\omega\in\mathscr{S}$. The equation $Lz_k=\omega(k)$ is the *recurrence relation* for the sequence $\{z_k\}\in\mathscr{S}$; the *order of the recurrence relation* is the order of the difference operator L.

3. Properties of the Jacobi coefficients. The well-known difference properties of the Jacobi polynomials,

$$\begin{aligned}(2k+\lambda-1)_3 xP_k^{(\alpha,\beta)}(x)&=2(k+\alpha)(k+\beta)(2k+\lambda+1)P_{k-1}^{(\alpha,\beta)}(x)\\&\quad+(\alpha^2-\beta^2)(2k+\lambda)P_k^{(\alpha,\beta)}(x)\\&\quad+2(k+1)(k+\lambda)(2k+\lambda-1)P_{k+1}^{(\alpha,\beta)}(x),\end{aligned}$$

$$\begin{aligned}2\Big[(k+\lambda-1)_2(2k+\lambda-1)\frac{d}{dx}P_{k+1}^{(\alpha,\beta)}(x)&+(\alpha-\beta)\\&\cdot(k+\lambda-1)(2k+\lambda)\frac{d}{dx}P_k^{(\alpha,\beta)}(x)\\&-(k+\alpha)(k+\beta)(2k+\lambda+1)\frac{d}{dx}P_{k-1}^{(\alpha,\beta)}(x)\Big]\\&=(k+\lambda-1)(2k+\lambda-1)_3P_k^{(\alpha,\beta)}(x)\end{aligned}$$

(see, e.g., [3, v. 2, §10.8] or [11, v. 1, §8]) imply the following basic identities for the Jacobi coefficients [9]:

$$Xb_k[f]=b_k[xf(x)], \tag{3.1}$$

$$Db_k[f']=b_k[f]. \tag{3.2}$$

Here X and D are difference operators,

$$X=\sum_{j=0}^{2}\xi_j(k)E^{j-1}, \tag{3.3}$$

$$D=\sum_{j=0}^{2}\delta_j(k)E^{j-1}, \tag{3.4}$$

and ξ_j, δ_j are rational functions,

$$\begin{aligned}&\delta_0(k)=2(k+\alpha)(2k+\lambda-3)/\gamma(k), \qquad \delta_1(k)=2(\alpha-\beta)(2k+\lambda)/\gamma(k),\\ &\delta_2(k)=-2(k+\beta+1)(2k+\lambda+3)/\gamma(k), \qquad \gamma(k)=(2k+\lambda-1)_3,\end{aligned} \tag{3.5}$$

$$\xi_0(k)=k\delta_0(k), \quad \xi_1(k)=\frac{1-\lambda}{2}\delta_1(k), \quad \xi_2(k)=-(k+\lambda)\delta_2(k). \tag{3.6}$$

From (3.1) and (3.2) we deduce the more general equations

$$p(X)b_k[f]=b_k[pf] \qquad (p \text{ a polynomial}), \tag{3.7}$$

$$D^ib_k[f^{(i)}]=b_k[f]. \tag{3.8}$$

Powers of the operators X and D may be obtained by the following procedure. Let $\Lambda\in\mathscr{L}$, $\Lambda=\sum_{j=0}^{2}\lambda_j(k)E^{j-1}$. Then we have

$$\Lambda^i=\sum_{j=0}^{2i}\lambda_{ij}(k)E^{j-i} \qquad (i\geqq 0),$$

where $\lambda_{ij}\in\mathscr{S}_{\text{rat}}$, $\lambda_{00}(k)\equiv 1$, and

$$\lambda_{ij}(k)=\lambda_0(k)\lambda_{i-1,j}(k-1)+\lambda_1(k)\lambda_{i-1,j-1}(k)+\lambda_2(k)\lambda_{i-1,j-2}(k+1)$$
$$(j=0,1,\cdots,2i;\lambda_{i,-2}=\lambda_{i,-1}=\lambda_{i,2i+1}=\lambda_{i,2i+2}\equiv 0)$$

for $i\geqq 1$. Moreover, if Λ has the *symmetry property* $\lambda_j(k)=\lambda_{2-j}(-k-\lambda)(j=0,1,2)$ then Λ^i also has this property: $\lambda_{ij}(k)=\lambda_{i,2i-j}(-k-\lambda)(j=0,1,\cdots,2i)$. Note that D and X have the symmetry property.

DEFINITION 3.1. Let

$$A_m^{(\varepsilon)}=I+\tau_m^{(\varepsilon)}(k)E,$$
$$Q_m^{(\varepsilon)}=kI-(k+\lambda+m+1)\tau_m^{(\varepsilon)}(k)E,$$

where $\varepsilon=\pm 1$, $m=0,1,\cdots$, and

$$\tau_m^{(-1)}(k)=\frac{(2k+\lambda+1)_3}{(2k+\lambda+m+1)_2(2k+\lambda-1)}, \qquad \tau_m^{(1)}(k)=-\frac{k+\beta+1}{k+\alpha+1}\tau_m^{(-1)}(k).$$

Further, let

$$\begin{aligned} S_{ij}^{(\varepsilon)} &= I && (i<j),\\ &= A_i^{(\varepsilon)} S_{i-1,j}^{(\varepsilon)} && (i \geqq j \geqq 0),\\ P_h^{(\varepsilon)} &= S_{h-1,0}^{(\varepsilon)} && (h \geqq 0),\\ R_m^{(\varepsilon)} &= I && (m=0),\\ &= Q_{m-1}^{(\varepsilon)} R_{m-1}^{(\varepsilon)} && (m \geqq 1). \end{aligned}$$

In virtue of [9, Lemma 3.5], we have

$$A_0^{(\varepsilon)}(X+\varepsilon I) = Q_0^{(\varepsilon)} D \qquad (\varepsilon = \pm 1). \tag{3.9}$$

It can be checked that

$$A_m^{(\varepsilon)} Q_{m-1}^{(\varepsilon)} = Q_m^{(\varepsilon)} A_{m-1}^{(\varepsilon)} \qquad (m \geqq 1). \tag{3.10}$$

LEMMA 3.1. *The identity*

$$P_i^{(\varepsilon)} b_k[\omega_\varepsilon^i f] = R_i^{(\varepsilon)} b_k[f], \tag{3.11}$$

where $\omega_\varepsilon = (x+\varepsilon)(d/dx)$, *holds for* $\varepsilon = \pm 1$ *and* $i = 0, 1, \cdots$.

Proof (by induction on i). For $i=0$ equation (3.11) is obviously true. Now, we have

$$P_1^{(\varepsilon)} b_k[\omega_\varepsilon f] = A_0^{(\varepsilon)}(X+\varepsilon I) b_k[f'] = Q_0^{(\varepsilon)} D b_k[f'] = R_1^{(\varepsilon)} b_k[f],$$

as can be seen using Definition 3.1, (3.9) and (3.4). Assume that (3.11) holds for a certain i ($i \geqq 1$). We obtain

$$P_{i+1}^{(\varepsilon)} b_k[\omega_\varepsilon^{i+1} f] = A_i^{(\varepsilon)} P_i^{(\varepsilon)} b_k[\omega_\varepsilon^i(\omega_\varepsilon f)] = A_i^{(\varepsilon)} R_i^{(\varepsilon)} b_k[\omega_\varepsilon f].$$

From (3.10) we deduce that

$$A_i^{(\varepsilon)} R_i^{(\varepsilon)} = Q_i^{(\varepsilon)} Q_{i-1}^{(\varepsilon)} \cdots Q_1^{(\varepsilon)} P_1^{(\varepsilon)},$$

which together with the result of the first part of the proof leads to

$$P_{i+1}^{(\varepsilon)} b_i[\omega^{i+1} f] = R_{i+1}^{(\varepsilon)} b_k[f]. \qquad \square$$

DEFINITION 3.2. Let

$$H_m^{(\varepsilon)} = H + m(X+\varepsilon I) \qquad (m = 0, 1, \cdots),$$

$$V_i^{(\varepsilon)} = \begin{cases} I & (i=0), \\ V_{i-1}^{(\varepsilon)} H_i^{(\varepsilon)} & (i = 1, 2, \cdots), \end{cases}$$

where $\varepsilon = \pm 1$ and

$$H = (k-1)_2 \delta_0(k) E^{-1} - k(k+\lambda)\delta_1(k) I + (k+\lambda)_2 \delta_2(k) E.$$

LEMMA 3.2. *For every* $m = 0, 1, \cdots$ *and* $\varepsilon = \pm 1$ *we have*

$$b_k[\omega_\varepsilon^m \{(x-\varepsilon)^m f(x)\}] = V_m^{(\varepsilon)} b_k[f]. \tag{3.12}$$

Proof. It can be checked (see [9]) that

$$b_k[(x^2-1) f'(x)] = H b_k[f]. \tag{3.13}$$

We also need the identity

$$\omega_\varepsilon\{(x-\varepsilon)^{m+1}f(x)\}=(x-\varepsilon)^m\{(x^2-1)f'(x)+(m+1)(x+\varepsilon)f(x)\} \tag{3.14}$$
$$(m=0,1,\cdots).$$

Equation (3.12) holds trivially for $m=0$. Assuming that it is true for a certain m ($m\geqq 0$), and making use of (3.14) and (3.13), we obtain

$$b_k\left[\omega_\varepsilon^{m+1}\{(x-\varepsilon)^{m+1}f(x)\}\right]=V_m^{(\varepsilon)}b_k\left[(x^2-1)f'(x)+(m+1)(x+\varepsilon)f(x)\right]$$
$$=V_m^{(\varepsilon)}\{H+(m+1)(X+\varepsilon I)\}b_k[f]=V_{m+1}^{(\varepsilon)}b_k[f]. \quad \square$$

4. Recurrence relation for the Jacobi coefficients. The main result of this paper is contained in Theorem 4.1 below. We shall need one more lemma.

LEMMA 4.1. *For every $r=0,1,\cdots$ and $\varepsilon=\pm 1$ we have*

$$\frac{d^r}{dx^r}\{(x+\varepsilon)^r f(x)\}=\sum_{h=0}^{r}\beta_h^{(r)}\omega_\varepsilon^{r-h}f(x), \tag{4.1}$$

where $\omega_\varepsilon=(x+\varepsilon)(d/dx)$,

$$\beta_h^{(r)}=(-1)^h\binom{r}{h}B_h^{(r+1)} \qquad (h=0,1,\cdots,r), \tag{4.2}$$

and $B_m^{(a)}$ are generalized Bernoulli numbers defined implicitly by

$$\left(\frac{t}{e^t-1}\right)^a=\sum_{m=0}^{\infty}\frac{t^m}{m!}B_m^{(a)}.$$

Proof. We start with the identity [11, v. 1, Eq. 2.8(16)]

$$(y+1)_r=\sum_{h=0}^{r}\beta_h^{(r)}y^{r-h}.$$

Combining this with [11, v. 1, Eq. 2.9(4)]

$$\frac{d^r}{dz^r}\{z^r g(z)\}=\prod_{i=1}^{r}(\omega+i)g(z),$$

where $\omega=z(d/dz)$, we obtain

$$\frac{d^r}{dz^r}\{z^r g(z)\}=\sum_{h=0}^{r}\beta_h^{(r)}\omega^{r-h}g(z). \tag{4.3}$$

For $\varepsilon\in\{-1,1\}$, let $z=x+\varepsilon$ and $f(x)=g(x+\varepsilon)$. Obviously, $\omega_\varepsilon f(x)=\omega g(z)$ and (4.1) is simply a transcription of (4.3). $\square$

The quantities (4.2) can be calculated recursively using formulae (cf. [11, v. 1, Eq. 2.8(7)])

$$\begin{aligned}&\beta_h^{(r)}=\beta_h^{(r-1)}+r\beta_{h-1}^{(r-1)} \qquad (r=1,2,\cdots;h=0,1,\cdots),\\ &\beta_0^{(r)}=1, \quad \beta_{r+1}^{(r)}=0 \qquad (r\geqq 0).\end{aligned} \tag{4.4}$$

THEOREM 4.1. *Let f be a function satisfying the differential equation*

$$\sum_{m=0}^{n} (q_m f)^{(m)} = q \tag{4.5}$$

of order n, where $q_0, q_1, \cdots, q_n$ are polynomials, and assume that $f^{(n)}$ can be expanded into a uniformly convergent Jacobi series. Let e_{im} be an integer $\geqq 0$ such that the equation

$$q_m(x) = (x+1)^{e_{1m}}(x-1)^{e_{-1m}} u_m(x) \qquad (m = 1, 2, \cdots, n; q_m \not\equiv 0) \tag{4.6}$$

holds for a polynomial u_m, $u_m(\pm 1) \neq 0$; let

$$s_i = \max\left\{ \max_{1 \leqq m \leqq n,\, q_m \not\equiv 0} (m - e_{im}), 0 \right\} \qquad (i = \pm 1), \tag{4.7}$$

and let $\varepsilon \in \{-1, 1\}, s, \sigma$ and d be integers,

$$\varepsilon = \begin{cases} 1 & \text{for } s_1 \leqq s_{-1}, \\ -1 & \text{for } s_1 > s_{-1}, \end{cases} \tag{4.8}$$

$$s = s_\varepsilon, \quad \sigma = s_{-\varepsilon}, \quad d = \sigma - s. \tag{4.9}$$

Finally, define the polynomials

$$\phi_h(x) = \sum_{i=s+h}^{n} \beta_{i-s-h}^{(i-s)} (x+\varepsilon)^{s-i} q_i(x) \quad \text{for } h = 0, 1, \cdots, n-s, \tag{4.10}$$

$$\psi_j(x) = (x-\varepsilon)^{-j} \phi_{j+d}(x) \quad \text{for } j = 1, 2, \cdots, n-\sigma. \tag{4.11}$$

Then we have the recurrence relation

$$Lb_k[f] = \rho(k), \tag{4.12}$$

where

$$L = P_d^{(\varepsilon)} \sum_{m=0}^{s-1} D^{s-m} q_m(X) + \sum_{h=0}^{d} S_{d-1,h}^{(\varepsilon)} R_h^{(\varepsilon)} \phi_h(X) + R_d^{(\varepsilon)} \sum_{j=1}^{n-\sigma} V_j^{(\varepsilon)} \psi_j(X), \tag{4.13}$$

and

$$\rho(k) = P_d^{(\varepsilon)} D^s b_k[q]. \tag{4.14}$$

(The notation used is that of Definitions 3.1 *and* 3.2.*) The order of this relation is expressed by the formula*

$$r = s_{-1} + s_1 + 2 \max_{0 \leqq m \leqq n,\, q_m \not\equiv 0} (\deg(q_m) - m). \tag{4.15}$$

Proof. Equation (4.5) implies that

$$\sum_{m=0}^{n} b_k\left[(q_m f)^{(m)}\right] = b_k[q].$$

Applying the operator D^s, where s is defined in (4.9), to both sides of the above equation, and using (3.7) and (3.8), we obtain

$$\sum_{m=0}^{s-1} D^{s-m} q_m(X) b_k[f] + \sum_{m=s}^{n} b_k\left[(q_m f)^{(m-s)}\right] = D^s b_k[q]. \tag{4.16}$$

Now, it readily follows from (4.7)–(4.9) that $e_{\varepsilon m} \geqq m-s \geqq m-\sigma$. Hence the formulae (4.10) and (4.11) actually define polynomials. Let $v_m(x)=(x+\varepsilon)^{s-m} q_m(x)$ for $m=s, s+1, \cdots, n$. Using Lemma 4.1 we transform the second sum on the l.h.s. of (4.16):

$$\begin{aligned}\sum_{m=s}^{n} b_k\left[(q_m f)^{(m-s)}\right] &= \sum_{r=0}^{n-s} b_k\left[\left\{(x+\varepsilon)^r v_{s+r}(x) f(x)\right\}^{(r)}\right] \\ &= \sum_{r=0}^{n-s} \sum_{h=0}^{r} \beta_h^{(r)} b_k\left[\omega_\varepsilon^{r-h}(v_{s+r} f)\right] = \sum_{h=0}^{n-s} b_i\left[\omega_\varepsilon^h(\phi_h f)\right].\end{aligned}$$

In the last expression, ϕ_h is the polynomial defined in (4.10).

Let ε and d be the integers defined in (4.8) and (4.9), respectively. Applying the operator $P_d^{(\varepsilon)}$ to both sides of (4.16) and making use of Lemma 3.1 and of (3.7), we obtain

$$\begin{aligned}&\left\{P_d^{(\varepsilon)} \sum_{m=0}^{s-1} D^{s-m} q_m(X) + \sum_{h=0}^{d} S_{d-1,h}^{(\varepsilon)} R_h^{(\varepsilon)} \phi_h(X)\right\} b_k[f] \\ &\qquad + R_d^{(\varepsilon)} \sum_{h=d+1}^{n-s} b_k\left[\omega_\varepsilon^{h-d}(\phi_h f)\right] = P_d^{(\varepsilon)} D^s b_k[q].\end{aligned} \tag{4.17}$$

Using (4.11) and Lemma 3.2, we deduce that

$$\sum_{h=d+1}^{n-s} b_k\left[\omega_\varepsilon^{h-d}(\phi_h f)\right] = \sum_{j=1}^{n-\sigma} b_k\left[\omega_\varepsilon^j\left\{(x-\varepsilon)^j \psi_j(x) f(x)\right\}\right] = \sum_{j=1}^{n-\sigma} V_j^{(\varepsilon)} \psi_j(X) b_k[f].$$

Noting this result in (4.17), equation (4.12) follows, in which the operator $L \in \mathscr{L}$ and the function $\rho \in \mathscr{S}$ are given by (4.13) and (4.14), respectively.

As we remarked in §1, the differential equation (4.5) is equivalent to the equation

$$\sum_{m=0}^{n} p_m f^{(m)} = q$$

of order n, where p_m are polynomials. Now, we have seen that the recurrence relation (4.12) is obtained by the use of the operator $P = P_d^{(\varepsilon)} D^s$ which satisfies

$$P \sum_{m=0}^{n} b_k\left[p_m f^{(m)}\right] = L b_k[f],$$

where L is the operator (4.13). According to [9, Lemma 4.2], the order of the operator L is equal to

$$r = r(P) + 2 \max_{0 \leqq m \leqq n,\, p_m \not\equiv 0} (\deg(p_m) - m).$$

It can be checked that, without affecting its validity, we may replace p_m by q_m in the above expression. As $r(P)=r(P_d^{(\varepsilon)})+r(D^s)=d+2s=s+\sigma$, we obtain the formula (4.15). □

The recurrence relation (4.12) takes a particularly simple form in the case where neither $x+1$ nor $x-1$ divides the coefficient $q_n(x)$ in the equation (4.5). Namely, equations (4.13) and (4.14) then become

$$L=\sum_{m=0}^{n} D^{n-m}q_m(X) \tag{4.18}$$

and

$$\rho(k)=D^n b_k[q],$$

respectively. The operator (4.18) has the order

$$2n+2\Big(\max_{0\leqq m\leqq n,\, q_m\not\equiv 0}(\deg(q_m)-m)\Big). \tag{4.19}$$

In the case under consideration our method is equivalent to the Paszkowski-type method described in [9, §5], which like most other algorithms [4–7], [12–15]) does *not* analyse the form of the coefficients of (4.5) but leads "blindly" to a recurrence relation of the *maximum* order (4.19). (Obviously, (4.19) is the upper bound for (4.15).)

There are two other special cases. First, if $q_m(x)=(x^2-1)^m w_m(x)$, where w_m is a polynomial ($m=0,1,\cdots,n$), then (4.13)–(4.15) reduce to

$$\begin{aligned} L&=\sum_{j=0}^{n} V_j^{(1)}\psi_j(X),\\ \rho(k)&=b_k[q],\\ r&=2\max_{0\leqq m\leqq n,\, w_n\not\equiv 0}(\deg(w_m)+m), \end{aligned}$$

where

$$\psi_j(x)=\sum_{m=j}^{n}\beta_{m-j}^{(m)}(x-1)^{m-j}w_m(x).$$

Second, when for $\varepsilon\in\{-1,1\}$, $q_m(x)=(x+\varepsilon)^m v_m(x)$, where v_m is a polynomial ($m=0,1,\cdots,n$) and $q_n(-\varepsilon)\neq 0$, we have

$$\begin{aligned} L&=\sum_{h=0}^{n} S_{n-1,h}^{(\varepsilon)}R_h^{(\varepsilon)}\phi_h(X),\\ \rho(k)&=P_n^{(\varepsilon)}b_k[q],\\ r&=n+2\max_{0\leqq m\leqq n,\, v_m\not\equiv 0}\deg(v_m), \end{aligned}$$

where

$$\phi_h(x)=\sum_{m=h}^{n}\beta_{m-h}^{(m)}v_m(x)\qquad(h=0,1,\cdots,n).$$

Finally, a symmetry property of the equation (4.12) should be noted which seems to be useful for checking purposes.

THEOREM 4.2. *Operator* (4.13) *can be written in the form*

$$L=\mu_d^{(\varepsilon)}(k)\sum_{i=0}^{r}\lambda_i(k)E^{i-u} \qquad (r=r(L))$$

in which

$$u=s+\max_{0\leqq m\leqq n,\, q_m\not\equiv 0}\bigl(\deg(q_m)-m\bigr),$$

$$\mu_d^{(-1)}(k)=(k+\alpha+1)_d\mu_d^{(1)}=\begin{cases}1 & (d=0),\\ (2k+\lambda-1)^{-1} & (d=1),\\ \bigl[(2k+\lambda-1)(2k+\lambda+d+1)_{d-2}\bigr]^{-1} & (d>1),\end{cases}$$

and $\lambda_0, \lambda_1, \cdots, \lambda_r \in \mathscr{S}_{\text{rat}}$ *are such that*

$$\lambda_{r-i}(k)=(-1)^{\eta}\lambda_i(-k-\lambda-d) \qquad (i=0,1,\cdots,r),$$

where

$$\eta=\begin{cases}d & (d\leqq 1),\\ d-1 & (d>1).\end{cases}$$

Proof. We give a short sketch of the proof. Let P represent any of the following operators:

$$P_d^{(\varepsilon)}, \quad R_d^{(\varepsilon)}, \quad S_{d-1,h}^{(\varepsilon)}R_h^{(\varepsilon)} \qquad (h=0,1,\cdots,d).$$

It can be shown by induction on d that

$$P=\mu_d^{(\varepsilon)}(k)\sum_{i=0}^{d}\pi_i(k)E^i$$

in which $\pi_0, \pi_1, \cdots, \pi_d \in \mathscr{S}_{\text{rat}}$ are such that

$$\pi_{d-i}(k)=(-1)^{\eta}\pi_i(-k-\lambda-d) \qquad (i=0,1,\cdots,d).$$

Further, let T stand for any of the operators

$$\sum_{m=0}^{s-1}D^{s-m}q_m(X), \quad \sum_{j=1}^{n-\sigma}V_j^{(\varepsilon)}\psi_j(X), \quad \phi_h(X) \qquad (h=0,1,\cdots,d).$$

From a symmetry property of the operators $V_j^{(\varepsilon)}$ and of the powers of D and X, it can be seen that

$$T=\sum_{j=0}^{2t}\tau_j(k)E^{j-t}$$

in which

$$t=s+\max_{m\in M,\, q_m\not\equiv 0}\bigl(\deg(q_m)-m\bigr),$$

and the coefficients τ_j satisfy

$$\tau_{2t-j}(k)=\tau_j(-k-\lambda) \qquad (j=0,1,\cdots,2l).$$

Here M is the set $\{0,1,\cdots,s-1\}$, or $\{\sigma+1,\sigma+2,\cdots,n\}$, or $\{s+h,s+h+1,\cdots,n\}$ $(h=0,1,\cdots,d)$, respectively.

Now, it can be checked that

$$PT=\mu_d^{(\varepsilon)}(k)\sum_{l=0}^{d+2t}\chi_l(k)E^{l-t}$$

in which χ_l are such that

$$\chi_{d+2t-l}(k)=(-1)^{\eta}\chi_l(-k-\lambda-d)\qquad(l=0,1,\cdots,d+2l).$$

Substituting this in (4.13) gives the theorem. □

5. Recurrence relation for the Gegenbauer coefficients. The special case $\alpha=\beta$ of the Jacobi series (1.1) is of particular importance. As has already been remarked in §1, it is then more convenient to deal with the Gegenbauer series (1.9) or the Chebyshev series (1.5). A recurrence relation for the Gegenbauer coefficients (1.16) can be constructed by a method analogous to that used in §4. However, in the present section we obtain neater looking results.

First of all, the basic identities (3.1), (3.2) may be replaced by

$$\bar{X}c_k[f]=c_k[xf(x)],\tag{5.1}$$

$$\bar{D}c_k[f']=c_k[f],\tag{5.2}$$

in which

$$\bar{X}=2^{-1}(k+\nu)^{-1}\{kE^{-1}+(k+2\nu)E\},$$
$$\bar{D}=2^{-1}(k+\nu)^{-1}\{E^{-1}-E\},\qquad \nu=\alpha+\tfrac{1}{2}.$$

We have

$$p(\bar{X})c_k[f]=c_k[pf]\qquad(p\text{ a polynomial}),\tag{5.3}$$

$$\bar{D}^ic_k[f^{(i)}]=c_k[f].\tag{5.4}$$

Note that [8]

$$\bar{D}^i=2^{-i}(k+\nu-i)_{2i+1}^{-1}\sum_{m=0}^{i}\rho_{im}(k+\nu)E^{2m-i},\tag{5.5}$$

where

$$\rho_{im}(\kappa)=(-1)^m\binom{i}{m}(\kappa-i)_m(\kappa-i+2m)(\kappa+m+1)_{i-m}\qquad(m=0,1,\cdots,i).$$

Further,

$$\bar{X}^i=2^{-i}\sum_{j=0}^{i}\xi_{ij}(k+\nu)E^{2j-i},$$

where

$$\xi_{ij}(k)=\binom{i}{j}\qquad(j=0,1,\cdots,i)$$

for $\nu=0$ and

$$\xi_{00}(\kappa)=1,$$
$$\xi_{ij}(\kappa)=\kappa^{-1}\{(\kappa-\nu)\xi_{i-1,j}(\kappa-1)+(\kappa+\nu)\xi_{i-1,j-1}(\kappa+1)\}$$
$$(j=0,1,\cdots,i;\, i\geqq 1;\, \xi_{i-1,-1}=\xi_{i-1,i}=0)$$

for $\nu\neq 0$ (ibid.).

DEFINITION 5.1. For any $m=0,1,\cdots$ and $\varepsilon=\pm 1$ we define

$$\overline{A}_m^{(\varepsilon)}=I-\varepsilon\tau_m(k)E,$$
$$\overline{Q}_m^{(\varepsilon)}=kI+\varepsilon(k+2\nu+m+1)\tau_m(k)E,$$

where

$$\tau_m(k)=(2k+2\nu+1)_2/(2k+2\nu+m+1)_2.$$

Further, let

$$\overline{S}_{ij}^{(\varepsilon)}=\begin{cases} I & (i<j),\\ \overline{A}_i^{(\varepsilon)}\overline{S}_{i-1,j}^{(\varepsilon)} & (i\geqq j\geqq 0),\end{cases}$$
$$\overline{P}_m^{(\varepsilon)}=\overline{S}_{m-1,0}^{(\varepsilon)} \qquad (i\geqq 0),$$
$$\overline{R}_i^{(\varepsilon)}=\begin{cases} I & (i=0),\\ \overline{Q}_i^{(\varepsilon)}\overline{R}_{i-1}^{(\varepsilon)} & (i>0).\end{cases}$$

Finally, let

$$\overline{H}_m^{(\varepsilon)}=\overline{H}+m(\overline{X}+\varepsilon I) \qquad (m=1,2,\cdots;\varepsilon=\pm 1),$$
$$\overline{V}_i^{(\varepsilon)}=\begin{cases} I & (i=0),\\ \overline{V}_{i-1}^{(\varepsilon)}\overline{H}_i^{(\varepsilon)} & (i\geqq 1),\end{cases}$$

where

$$\overline{H}=2^{-1}(k+\nu)^{-1}\{(k-1)_2E^{-1}-(k+2\nu)_2E\}.$$

Lemmata 3.1 and 3.2 now have the following analogues.

LEMMA 5.1. *For any $i\geqq 0$ and $\varepsilon=\pm 1$ we have*

$$\overline{P}_i^{(\varepsilon)}c_k[\omega_\varepsilon^i f]=\overline{R}_i^{(\varepsilon)}c_k[f],$$

where $\omega_\varepsilon=(x+\varepsilon)(d/dx)$.

LEMMA 5.2. *For any $j\geqq 0$ and $\varepsilon=\pm 1$ we have the identity*

$$c_k[\omega_\varepsilon^j\{(x-\varepsilon)^j f(x)\}]=\overline{V}_j^{(\varepsilon)}c_k[f].$$

Finally, we have the following

THEOREM 5.1. *Let f be a function satisfying the differential equation* (4.5) *of order n and such that its nth derivative can be expanded into the uniformly convergent Gegenbauer series. We then have the recurrence relation*

$$\overline{L}c_k[f]=\overline{\rho}(k), \tag{5.6}$$

where

$$\overline{L}=\overline{P}_d^{(\varepsilon)}\sum_{m=0}^{s-1}\overline{D}^{s-m}q_m(\overline{X})+\sum_{h=0}^{d}\overline{S}_{d-1,h}^{(\varepsilon)}\overline{R}_h^{(\varepsilon)}\phi(\overline{X})+\overline{R}_d^{(\varepsilon)}\sum_{j=1}^{n-\sigma}\overline{V}_j^{(\varepsilon)}\psi_j(\overline{X}),$$

$$\overline{\rho}(k)=\overline{P}_d^{(\varepsilon)}\overline{D}^s c_k[q],$$

and the symbols $\varepsilon, s, \sigma, d, \phi_h, \psi_j$ *have the meaning given in Theorem* 4.1. *The order of the relation* (5.6) *is expressed by* (4.15). *Operator* $\overline{L}$ *can be written in the form*

$$\overline{L}=\mu(k)\sum_{i=0}^{r}\lambda_i(k)E^{i-u}, \qquad r=r(L),$$

in which

$$u=s+\max_{0\leqq m\leqq n,\, q_m\not\equiv 0}(\deg(q_m)-m),$$

$$\mu(k)=\begin{cases}1 & (d=0),\\ 1/(2k+2\nu+d+1)_{d-1} & (d>0),\end{cases}$$

and $\lambda, \lambda_1, \cdots, \lambda_r \in \mathscr{S}_{\mathrm{rat}}$ *are such that*

$$\lambda_{r-i}(k)=\eta\varepsilon^d\lambda_i(-k-2\nu-d) \qquad (i=0,1,\cdots,r),$$

where $\eta=0$ *for* $d=0$ *and* $\eta=-1$ *for* $d>0$.

From (4.5), we have the identity

$$\sum_{m=0}^{n}c_k\left[(q_m f)^{(m)}\right]=c_k[q].$$

Applying the operator $\overline{D}^n$ to both sides of the above equation and using (5.3) and (5.4), we obtain another recurrence relation obeyed by $\{c_k[f]\}$, namely

$$\left\{\sum_{m=0}^{n}\overline{D}^{n-m}q_m(\overline{X})\right\}c_k[f]=\overline{D}^n c_k[q] \tag{5.7}$$

or, in view of (5.5),

$$\left\{\sum_{m=0}^{n}2^m(k+\nu-n)_m(k+\nu-m+1)_m\sum_{i=0}^{n-m}\rho_{n-m,i}(k+\nu)E^{2i-n+m}q_m(\overline{X})\right\}c_k[f]$$

$$=\sum_{m=0}^{n}\rho_{nm}(k+\nu)c_{k-n+2m}[q].$$

The described procedure is equivalent to the generalized Paszkowski algorithm ([8]; or [17, p. 196 ff]). Clearly, equation (5.7) has order (4.19). Equations (5.6) and (5.7) coincide if and only if the coefficient q_n in (4.5) has no linear factors of the form $(x+1)$ or $(x-1)$.

6. Example. The generalized hypergeometric function

$$F(z)={}_{p+1}F_p\left(\begin{matrix}\gamma_1,\gamma_2,\cdots,\gamma_{p+1}\\ \varepsilon_1,\varepsilon_2,\cdots,\varepsilon_p\end{matrix}\middle|z\right) \qquad \left(\mathrm{Re}\left(\sum\gamma_i-\sum\varepsilon_i\right)<0\right) \tag{6.1}$$

satisfies the differential equation [3, v. 1, §4.2], or [11, v. 1, §5.1]

$$\left\{ z \prod_{i=1}^{p+1} (\omega+\gamma_i) - \omega \prod_{j=1}^{p} (\omega+\varepsilon_j-1) \right\} Y(z)=0 \qquad \left(\omega = z\frac{d}{dz}\right)$$

which can also be written in the form

$$u_0 Y(z) + \sum_{m=1}^{p} (u_m z - v_m) z^{m-1} Y^{(m)}(z) + (z-1) z^p Y^{(p+1)}(z) = 0, \tag{6.2}$$

where u_m, v_m are constants. Therefore, the function

$$f_p(x) = F\left(\frac{1+x}{2}\right) \qquad (-1 \leqq x \leqq 1) \tag{6.3}$$

satisfies

$$u_0 y(x) + \sum_{m=1}^{p} (u_m x - w_m)(x+1)^{m-1} y^{(m)}(x) + (x-1)(x+1)^p y^{(p+1)}(x) = 0, \tag{6.4}$$

where $w_m = u_m - 2v_m$ $(m=1,2,\cdots,p)$.

It is known [11, v. 2, §9] that the function (6.3) can be expanded in the Jacobi series (1.1) and that

$$a^{(\alpha,\beta)}[f_p] = \frac{\prod_{i=1}^{p+1} (\gamma_i)_k}{(k+\lambda)_k \prod_{j=1}^{p} (\varepsilon_j)_k} p+2F p+1\left(\begin{matrix} k+\beta+1, k+\gamma_1, \cdots, k+\gamma_{p+1} \\ 2k+\lambda+1, k+\varepsilon_1, \cdots, k+\varepsilon_p \end{matrix} \middle| 1 \right). \tag{6.5}$$

We need the following alternative form of (6.4):

$$\lambda_0 y + \sum_{m=1}^{p+1} \left\{ (\lambda_m x + \mu_m)(x+1)^{m-1} y \right\}^{(m)} = 0,$$

where $\lambda_{p+1} = -\mu_{p+1} = 1$ (cf. (1.11), (1.12)). By virtue of Theorem 4.1 we have the recurrence relation

$$L_p b_k^{(\alpha,\beta)}[f_p] = 0$$

of order $p+1$ satisfied by the Jacobi coefficients of the function (6.3), where

$$L_p = \lambda_0 P_{p-1}^{(1)} D + \sum_{h=0}^{p-1} S_{p-2,h}^{(1)} R_h^{(1)} (\eta_h + X + \theta_h I) + R_{p-1}^{(1)} H_1^{(1)},$$

and

$$\eta_h = \sum_{l=h+1}^{p+1} \beta_{l-h-1}^{(l-1)} \lambda_l, \qquad \theta_h = \sum_{l=h+1}^{p+1} \beta_{l-h-1}^{(l-1)} \mu_l.$$

In particular, the Jacobi coefficients of the function

$$f_2(x) = {}_3F_2\left(\begin{matrix} \gamma_1, \gamma_2, \gamma_3 \\ \varepsilon_1, \varepsilon_2 \end{matrix} \middle| \frac{1+x}{2} \right)$$

obey the third-order recurrence relation

$$L_2 b_k^{(\alpha,\beta)}[f_2] = 0, \tag{6.6}$$

where

$$L_2 = A_0^{(1)}\{\lambda_0 D + (\lambda_1+\lambda_2+2)X + (\mu_1+\mu_2-2)I\} + Q_0^{(1)}\{H + (\lambda_2+4)X + (\mu_2-2)I\},$$

and

$$\begin{aligned}
&\lambda_0 = \sigma_1 - \sigma_2 + \gamma_1\gamma_2\gamma_3 - 1, \qquad \lambda_1 = \sigma_2 - 3\sigma_1 + 7,\\
&\lambda_2 = \sigma_1 - 6, \qquad \mu_1 = \sigma_2 - 3\sigma_1 - 2\varepsilon_1\varepsilon_2 + 4\varepsilon_1 + 4\varepsilon_2 - 1,\\
&\mu_2 = \sigma_1 - 2\varepsilon_1 - 2\varepsilon_2 + 1, \qquad \sigma = \gamma_1+\gamma_2+\gamma_3,\\
&\sigma_2 = \gamma_1\gamma_2 + \gamma_1\gamma_3 + \gamma_2\gamma_3.
\end{aligned}$$

Remark that an equivalent result is furnished by the optimum method of [9]. On the other hand, we have recently shown [10] that the quantities (6.5) satisfy the recurrence relation

$$\sum_{i=0}^{p+1} A_i(k)\, a_{k+i}^{(\alpha,\beta)}[f_p] = 0 \tag{6.7}$$

of order $p+1$, where the coefficients $A_i(k)$ are expressed in terms of hypergeometric functions of unit argument. We have checked that, for $p=2$, (6.7) is equivalent to (6.6).

Acknowledgment. The author would like to thank Professors S. Paszkowski and M. J. D. Powell for helpful comments.

REFERENCES

[1] C. W. CLENSHAW, *The numerical solution of linear differential equations in Chebyshev series*, Proc. Cambridge Phil. Soc., 53 (1957), pp. 134–139.

[2] D. ELLIOTT, *The expansion of functions in ultraspherical polynomials*, J. Austral. Math. Soc., 1 (1959–1960), pp. 428–438.

[3] A. ERDÉLYI et al., *Higher Transcendental Functions*, McGraw-Hill, New York, 1953.

[4] L. FOX, *Chebyshev methods for ordinary differential equations*, Comput. J., 4 (1962), pp. 318–331.

[5] L. FOX AND I. B. PARKER, *Chebyshev Polynomials in Numerical Analysis*, Oxford Univ. Press, London, England, 1968.

[6] K. O. GEDDES, *Symbolic computation of recurrence equations for the Chebyshev series solution of linear* ODE's, Proc. 1977 MACSYMA Users' Conference, Univ. of California, Berkeley, CA, 1977, NASA CP-2012, pp. 405–423.

[7] T. S. HORNER, *Recurrence relations for the coefficients in the Chebyshev series solutions of ordinary differential equations*, Math. Comp., 35 (1980), pp. 893–905.

[8] S. LEWANOWICZ, *Construction of a recurrence relation of the lowest order for coefficients of the Gegenbauer series*, Zastos. Mat., 15 (1976), pp. 345–396.

[9] _____, *Construction of the lowest-order recurrence relation for the Jacobi coefficients*, Zastos. Mat., 17 (1983), pp. 655–675.

[10] _____, *Recurrence relations for hypergeometric functions of unit argument*, Math. Comp., 45 (1985), pp. 521–535.

[11] Y. L. LUKE, *The Special Functions and their Approximations*, Academic Press, New York, 1969.

[12] A. MAGNUS, *Application des récurrences au calcul d'une classe d'intégrales*, Rep. 71, Inst. Math. Pure Appl., Univ. de Louvain, 1974.

[13] A. G. MORRIS AND T. S. HORNER, *Chebyshev polynomials in the numerical solution of differential equations*, Math. Comp., 31 (1977), pp. 881–891.

[14] O. OLUREMI OLAOFE, *On the Tchebyschev method of solution of ordinary differential equations*,, J. Math. Anal. Appl., 60 (1977), pp. 1–7.

[15] S. PASZKOWSKI, *Zastosowania numeryczne wielomianów i szeregów Czebyszewa*, PWN, Warszawa, 1975.

[16] N. ROBERTSON, *An* ALTRAN *program for finding a recursion formula for the Gegenbauer coefficients of a function*, Nat. Res. Inst. for Math. Sci., Spec. Rep. SWISK 11, Pretoria, 1979.

[17] J. WIMP, *Computation with Recurrence Relations*, Pitman, Boston, MA, 1984.

SIAM J. MATH. ANAL.
Vol. 17, No. 5, September 1986

002

INVARIANT CURVES FOR MAPPINGS*

HAL L. SMITH†

Abstract. The main result of this paper concerns a smooth map T of a Banach space X into itself which has an unstable fixed point x_0. We prove that if the spectral radius λ_0 of the Frechet derivative of T at x_0 is an eigenvalue which exceeds one and appropriate additional assumptions hold, then there is a smooth invariant curve emanating from x_0 which might be called the "most unstable manifold" of x_0. The curve is parametrized by a smooth function satisfying a functional equation involving T and λ_0. This result is shown to be especially useful when the map T possesses certain monotonicity conditions. In this case, the curve can be shown to be monotone and to terminate on a stable fixed point of T.

Key words. unstable manifold of a fixed point, functional equation, smooth linearization

AMS(MOS) subject classifications. Primary 47H99, 47H07, 39B70

Introduction. In 1901, motivated by the work of Poincaré on the stability of periodic solutions of ordinary differential equations, T. Hadamard [4] published a short note in which he showed that the unstable manifold of a hyperbolic fixed point of a smooth diffeomorphism T of the plane can be obtained as the limit of a sequence of curves generated by successively applying T to a suitable initial curve. In fact, his result is more general. He assumed that

$$T(x,y)=(sx+F(x,y),s'y+\Phi(x,y)),$$

where F and Φ have Taylor developments about the origin beginning with quadratic terms, $s>1$, $|s'|<s$. He then shows that if C is a curve which can be parametrized by $y=y(x)$ with $|dy/dx|\leqq\alpha$, then $C_n\equiv T^n(C)$ converges to an invariant curve for the map T (T^n denotes the n-fold composition of T with itself: $T\circ T\circ\cdots\circ T, n$ times). Since s' is not assumed to satisfy $|s'|<1$, nor $s'\neq 1$, this invariant curve, in general, will lie in the unstable manifold of the fixed point; it might be called (part of) the "most unstable manifold" of the fixed point.

The purpose of this paper is to prove a general version of this result, obtaining a "most unstable" invariant curve emanating from an unstable fixed point for a smooth map. We will also find a parametrization of this invariant curve which effectively linearize the action of T on the curve. The following is a special case of our main theorem.

THEOREM. *Let T: $X\to X$ be a smooth map of a real Banach space X into itself having a fixed point x_0. Let $A\equiv DT(x_0)$ be the derivative of T at x_0 and assume that $Ae_0=\lambda_0 e_0$ where $e_0\notin\overline{R(A-\lambda_0 I)}$ and $\lambda_0>1$ is the spectral radius of A. Then there exists a unique smooth function y: $[0,\infty)\to X$, $y(t)=x_0+te_0+O(t^2)$ as $t\to 0$, with the following properties.*

I. $y(t)=T(y(\lambda_0^{-1}t))$ $t\geqq 0$.

II. $y(t)=\lim_{n\to\infty}T^n(x_0+\lambda_0^{-n}te_0)$.

III. *If T is a diffeomorphism then y is a one-to-one immersion ($y'(t)\neq 0$) and $y(t)$ is not a fixed point of T (or periodic point) for any $t>0$.*

*Received by the editors June 22, 1984, and in revised form February 14, 1985. This paper was presented at the International Conference on Qualitative Theory of Differential Equations, Edmonton, Alberta, Canada, June 1984.

†Department of Mathematics, Arizona State University, Tempe, Arizona 85287.

In II, T^n again denotes the n-fold composition $T \circ T \circ T \circ \cdots \circ T$, n times, of T with itself. We will use this notation throughout the rest of the paper without further comment.

If we let C be the curve in X parametrized by $y, C = \{y(t): t \geqq 0\}$, then the functional relation I can be interpreted in terms of the commuting diagram of maps

$$\begin{array}{ccc} C & \xrightarrow{T} & C \\ {\scriptstyle y}\uparrow & & \uparrow{\scriptstyle y} \\ [0,\infty) & \xrightarrow[t \to \lambda_0 t]{} & [0,\infty) \end{array}$$

In other words, the parametrization y of C effects a linearization of T on C. One may view the functional relations I in a different vein by setting $f(s) = y(\lambda_0^s)$, $-\infty < s < \infty$, yielding the functional equation $f(s) = T(f(s-1))$.

The finite-dimensional version of the theorem can be obtained from existing results. First observe that the existence of y: $[0,\infty) \to X$ in the theorem follows by continuation from the existence of y: $[0,\varepsilon) \to X$, $\varepsilon > 0$, satisfying I, II and III. Hence, the issue is the local existence of y for t near zero. By using [11, Thm. 5.1], one obtains the existence of a local C^2 unstable manifold for T, which by a C^2 change of variables can be taken to be a neighborhood of x_0 in an affine subspace of X. Then, a result of Hartman ([5, Exercise 8.2, p. 246]) can be applied to the restriction of T to the unstable manifold yielding a C^1 change of coordinates near x_0 on the unstable manifold so that in the new coordinates, T (restricted to unstable subspace) can be represented by its linear part. This result is easily seen to imply our theorem.

Our motivation for proving the above theorem stems from dynamical systems theory. One may imagine that T is the Poincaré map associated with a flow. The unstable fixed point x_0 may represent an equilibrium solution or periodic orbit for the flow. In certain situations, one may be able to conclude that the invariant curve C, emanating from x_0, asserted to exist by our theorem, is precompact in the phase space X. In such cases, one can assert (see Remark 2 following Theorem 1.1) that the limit set

$$\Lambda = \left\{ x: x = \lim_{n \to \infty} y(t_n), t_n \to \infty \right\}$$

is nonempty, compact, connected and an invariant set for T. It may contain fixed points of T, periodic points, a closed curve or more exotic strange attractors. Those familiar with dynamical systems theory will have a ready supply of examples. In recent studies of chaotic motion and strange attractors, a standard technique has been to locate saddle fixed points of suitable diffeomorphisms and to numerically approximate the unstable manifold. For example, in (see in particular, [3, Figs. 2.2.7 and 2.2.8]) it is conjectured that the strange attractor in the periodically perturbed Duffing equation is the closure of such an unstable manifold corresponding to the Poincaré map.

In a different direction, Pounder and Rogers [7] study the difference equation $(x_n, y_n) = T_a(x_{n-1}, y_{n-1})$ associated with the mapping $T_a(x,y) = (y, ay(1-x))$. This equation is equivalent to the "delayed" logistic equation $y_{n+1} = ay_n(1-y_{n-1})$. They obtain an existence proof of an invariant curve, $C_\infty(a)$, for T_a, emanating from the saddle $(0,0)$ for $a > 1$ by techniques very similar to ours. Their numerical calculations of $C_\infty(a)$ for various values of a are particularly interesting, (see especially [3, Figs. 2a–c and 5–12]) showing clearly the complex behavior to be expected of the invariant curve $C = C_\infty(a)$ of our theorem. In [7, Fig. 2a] the set Λ consists of another fixed point with

C joining the two fixed points, $1<a<\frac{5}{4}$. For larger values of $\frac{5}{4}<a<2$, Λ is the same fixed point but C spirals around this fixed point. For $2<a\leqq 2.20$, C wraps around a closed invariant curve for T, then $\Lambda=S^1$. For $a\simeq 2.27$, the curve C loops back to the origin infinitely many times and it appears clear that the outer loop of C is contained in Λ in this case.

We envision that an important application of our result will be to discrete dynamical systems generated by a monotone map T. A monotone map is one which preserves a partial ordering on the space X, induced, for example, by a cone in X. Such maps arise naturally as Poincaré maps for ordinary differential equations, for which the Kamke theorem [12] applies, the so-called competitive and cooperative systems of Hirsch [12], [13] (see also Selgrade [8] and [16], [17]) and for parabolic partial differential equations which generate monotone flows in suitable function spaces via maximum principle arguments (see [14], [15]). If x_0 is an unstable fixed point of a monotone map T, the spectral assumptions of our theorem can often be verified by the Perron–Frobenius theorem in finite dimensions or the Krein–Rutman theorem in infinite dimesions. In §2 we give sufficient conditions for the function y of our theorem to be monotone with respect to the usual ordering on R^+ and the partial ordering on X, in case T is monotone. This result leads to the dichotomy: either C is monotone and unbounded or C joins x_0 to a semi-stable fixed point, x_∞, of T and C is tangent at x_∞ to the eigenvector corresponding to the dominant eigenvalue of $DT(x_\infty)$. This result provides the key tool for our analysis of periodic competitive and cooperative systems in future publications [16], [17].

1. Main results. Before stating and proving our main results, we establish some notation. We will introduce various Banach spaces in what follows and will reserve $|\cdot|$ for the norm. Since we also will consider Banach spaces whose elements are Banach space-valued functions there is the potential for confusion over which norm is being used. In all cases we have attached subscripts to function-space norms in order to reduce the chance for confusion. If Z is a Banach space and $r>0$ we write $B_r(0)$ for the open ball about zero of radius r, $B_r(0)=\{z\in Z\colon |z|<r\}$, and $\overline{B_r(0)}$ for the closure of $B_r(0)$ in Z. If T is a map from a Banach space X to a Banach space Y, we call T a C^n map if it possesses n continuous derivatives at all points of its domain.

THEOREM 1.1. *Let X be a real, Banach space, $\Omega\subseteq X$ an open set and $T\colon \Omega\to X$ be a C^1-map with fixed point $x_0\in\Omega$. Let T be C^2 in a neighborhood of x_0. Let $A\equiv DT(x_0)$ have spectral radius λ_0 and assume*

(a) $\lambda_0>1$,

(b) $Ae_0=\lambda_0 e_0$, $e_0\notin\overline{R(A-\lambda_0 I)}$.

Then there exists t_0, $0<t_0\leqq\infty$ and a unique C^1 function $y\colon [0,t_0)\to X$ with the properties

(i) $y(t)=x_0+te_0+O(t^2)$ *as* $t\to 0$,

(ii) $y(t)=T(y(\lambda_0^{-1}t))$, $0\leqq t<t_0$,

(iii) $x_n(t)\equiv T^n(x_0+\lambda_0^{-n}te_0)$, $n=0,1,2,\cdots$ *satisfies* $x_n\to y$ *uniformly on compact sets and* $x_n'(t)\to y'(t)$ *on* $[0,t_0)$,

(iv) *either* $t_0=+\infty$ *or* t_0 *is maximal with the property that* $y(t)\in\Omega$, $0<t<\lambda_0^{-1}t_0$,

(v) *if T is a diffeomorphism than $y\colon [0,t_0)\to X$ is an injective immersion* ($y'(t)\neq 0$).

Before proceeding to the proof, it is convenient to make the following remarks.

Remark 1. Condition (b) of Theorem 1.1 may seem strange. It implies the existence of a decomposition of $X=\operatorname{span} e_0+Z$, where Z is a closed subspace of T, which reduces A ($AZ\subset Z$). This, in fact, is all we use in the proof. Condition (b) will hold, for example, if λ_0 is a simple pole of $(A-\lambda I)^{-1}$ and e_0 is any nonzero vector in $N(A-\lambda_0 I)$.

Remark 2. Assume $t_0 = +\infty$ and $C = \{y(t)\colon t \geqq 0\}$ lies in a compact subset of X, for example, X might be finite-dimensional and C bounded. Let $\Lambda = \{x\colon x = \lim_{n\to\infty} y(t_n), \{t_n\}_{n=1}^{\infty}$ an arbitrary sequence satisfying $\lim t_n = +\infty\}$. It is easily seen that Λ is a nonempty, compact, connected, invariant set for T.

Remark 3. When T is a diffeomorphism, the map y is one-to-one and $y(t)$ is not a periodic point of T for any $T > 0$. Both properties can fall spectacularly if T is not one-to-one. An interesting example is given by $T\colon R \to R$ defined by $x \to 3x - 4x^3$, which lies $T(0) = 0$ and $DT(0) = 3$. Corresponding to $e_0 = +1$, $y\colon [0, \infty) \to R$ of Theorem 1.1 is given by $y(t) = \sin t$ since $\sin t = 3\sin(t/3) - 4(\sin(t/3)^3$. The set Λ of limit points of $y(t)$ is the invariant set $[-1, 1]$ for T. It can be seen that $T|_{[-1,1]}\colon [-1,1] \to [-1,1]$ exhibits Li–Yorke chaos [10] with periodic points of every period, an uncountable set of a-periodic trajectories and an invariant probability measure (ergodic [18]) $(1/\pi)\,dx/\sqrt{1-x^2}$. All this follows from the observation that $h\colon [-1,1] \to [-1,1]$ defined by $h(x) = \sin(\pi x/2)$ is a homeomorphism providing a conjugacy between $T|_{[-1,1]}$ and the piecewise linear map $S\colon [-1,1] \to [-1,1]$ defined by

$$S = \begin{cases} -3t - 2, & -1 \leqq t < -\frac{1}{3}, \\ 3t, & -\frac{1}{3} \leqq t < \frac{1}{3}, \\ -3t + 2, & \frac{1}{3} \leqq t \leqq 1. \end{cases}$$

Note that if $t_0 = 2 \cdot 3^{-3}$ and $t_i = S^i(t_0)$, then $t_3 < t_0 < t_1 < t_2$, establishing Li–Yorke chaos [10]. It is clear that S preserves Lebesgue measure.

Clearly, in this example, explicit knowledge of the function y was of great help in determining the dynamics of the difference equation $x_{n+1} = T(x_n)$. It is interesting to speculate on whether the dynamics of other difference equations generated by one-dimensional maps might be illuminated by this approach, although it is clearly too much to hope that y can be explicitly found in general. If, for example, T is a simple polynomial map, for example, $T(x) = rx(1-x)$, fixing the origin with $T'(0) = r$, $r > 1$, it may be convenient to allow x to be complex: $T\colon C \to C$. It is then not difficult to see that there is an entire function $y\colon C \to C$ with an essential singularity at ∞, $y(z) = z + O(|z|^2)$ as $|z| \to 0$, such that $y(z) = T(y(r^{-1}z))$ (see Proposition 1.4 or [7]). Although y is not a homeomorphism in general, it is surjective and one has the following commutative diagram

$$\begin{array}{ccc} C & \xrightarrow{\quad} & C \\ {\scriptstyle y}\downarrow & {\scriptstyle z \to rz} & \downarrow{\scriptstyle y} \\ C & \xrightarrow[\quad T \quad]{} & C \end{array}$$

Hence the iterates $w_{n+1} = T(w_n)$, $w_0 = y(z_0)$ are $w_n = y(r^n z_n)$.

Remark 4. There are many different properties of a map T which will insure that $t_0 = +\infty$. We mention only one. If there exists a closed invariant set K for T, $K \subset \Omega$, $(T(K) \subset K)$ such that for some $\varepsilon > 0$, $x_0 + te_0 \in K$ for $0 \leqq t \leqq \varepsilon$ then $t_0 = +\infty$ and $y(t) \in K$ for $t \geqq 0$. In order to see this, observe that $y(t) \in K$ by (iii) for $[0, t_0)$. But then, by (iv), $t_0 = +\infty$.

Remark 5. If T is a diffeomorphism, the hypotheses of Theorem 1.1 may hold for T^{-1} if T has a smallest eigenvalue. More precisely, if $\lambda_0 \in (0,1)$ is a simple eigenvalue of $DT(x_0)$ with corresponding eigenvector e_0 and $\inf\{|\lambda|\colon \lambda \in \sigma(DT(x_0)) - \{\lambda_0\}\} \geqq \lambda_0$ then one can apply Theorem 1.1. to T^{-1} to obtain the existence of a map $y\colon [0, t_0) \to X$ such that $y(\lambda_0 t) = T(y(t))$.

Remark 6. The smoothness assumptions on T of Theorem 1.1 may be relaxed. We use only that $DT(x)$ is uniformly Lipschitz continuous in a neighborhood of x_0. If T is only continuous on X, differentiable at x_0, and $r(x)=T(x)-x_0-A(x-x_0)$ satisfies (1.6) of Proposition 1.3 below, then a continuous y: $[0,t_0)\to X$ exists satisfying (ii) and (iii) except that the convergence may not be uniform on compact sets if $\dim X=+\infty$. (C^1 functions are uniformly continuous on bounded sets whereas continuous functions are not generally uniformly continuous on bounded subsets of infinite-dimensional spaces.)

Before proceeding to the proof of Theorem 1.1 we present a simple example. Let T: $R^2\to R^2$ be given by $T(z)=z^2$ where we have identified R^2 with C. Consider the fixed point $z_0=1$ of T where $DT(1)=\left(\begin{smallmatrix}2 & 0\\ 0 & 2\end{smallmatrix}\right)$. For every $\theta\in[0,2\pi]$, $e_\theta\equiv e^{i\theta}$ is an eigenvalue and the corresponding function y_0: $[0,\infty]\to R^2$, the existence of which is asserted in Theorem 1.1, is given by $y_\theta(t)=\exp(e_\theta t)$. As θ runs over $[0,2\pi)$ the curves $C_\theta=\{y_\theta(t)\colon \geqq 0\}$ cover $R^2-\{0\}$.

We begin the proof of Theorem 1.1 by establishing the existence of $y(t)$ satisfying (i), (ii) and (iii) for small t.

LEMMA 1.2. *Assume the hypotheses of Theorem* 1.1. *Then there exist* $\tau>0$ *and a* C^1 *function* y: $[0,\tau]\to\Omega$ *such that* (i), (ii) *and* (iii) *of Theorem* 1.1 *hold.*

Proof. It follows from our assumptions on A that there exists $f_0\in N(A^*-\lambda_0 I)$, where A^*: $X^*\to X^*$ is the adjoint of A on the dual space X^* of X, such that $f_0(e_0)=1$. The projection P: $X\to X$ defined by $Px=f_0(x)e_0$ leads to a decomposition of X: $X=\operatorname{lin}\{e_0\}+Z$, $Z=(I-P)X$, where $\operatorname{lin}\{e_0\}$ is the linear span of e_0, which reduces A: $A=\lambda_0 I+B$, B: $Z\to Z$. We may choose an equivalent norm on Z so that the corresponding operator norm of B satisfies $\|B\|<\lambda_0^2$ (spectral radius of $B\leqq\lambda_0$). For $x\in X$, we write $x=ue_0+z$, $u\in R$, $z\in Z$ and $|x|=|u|+|z|$ where $|u|$ is the absolute value of u.

We assume w.l.o.g. that $x_0=0$. If $Tx_0=x_0$ then $\bar S(x)=T(x+x_0)-x_0$ satisfies $\bar S(0)=0$, $D\bar S(0)=A$.

Write, for $x=ue_0+z$

$$T(x)=T_1(u,z)e_0+T_2(u,z)$$

where T_1: $R\times Z$ and T_2: $R\times Z\to Z$ satisfy

$$T_1(u,z)=\lambda_0 u+r_1(u,z),\qquad T_2(u,z)=Bz+r_2(u,z)$$

and $r_i(0,0)=0$, $Dr_i(0,0)=0$. Our assumption that T is C^2 in a neighborhood of 0 implies that there exists C, $\delta>0$ such that

$$\sum_{i=1}^{2}\|Dr_i(u_1,z_1)=Dr_i(u_2,xz_2)\|\leqq C\left[|u_1-u_2|+|z_1-z_2|\right] \tag{1.1}$$

provided $|u_i|+|z_i|\leqq\delta$, $i=1,2$.

One easily verifies that (1.1) implies that

$$\sum_{i=1}^{2}|Dr_i(x_1)h_1-Dr_i(x_2)h_2|\leqq C\left[\max_i|x_i||h_1-h_2|+\max_i|h_i||x_1-x_2|\right], \tag{1.2}$$

$$x_i=(u_i,z_i)\quad h_i=(h_{i1},h_{i2})$$

provided $|x_i|\leqq\delta$, $i=1,2$.

We seek a function $y(t)=te_0+o(t)$ satisfying (ii). Setting

$$y(t)=(t+\eta(t))e_0+z(t),\ z(t)\in Z,$$

(ii) will hold if and only if

$$\eta(t)=\lambda_0\eta(\lambda_0^{-1}t)+r_1(\lambda_0^{-1}t+\eta(\lambda_0^{-1}t),z(\lambda_0^{-1}t)), \tag{1.3a}$$

$$z(t)=Bz(\lambda_0^{-1}t)+r_2(\lambda_0^{-1}t+\eta(\lambda_0^{-1}t),z(\lambda_0^{-1}t)). \tag{1.3b}$$

We will solve (1.3) for (η,z) by the contraction mapping theorem.Let $\tau>0$ and let

$$C_\tau^1=\{x(t)=(\eta(t),z(t))\in C^1([0,\tau],X):$$
$$(\eta(0),z(0))=0,|\eta'(t)|+|z'(t)|\leqq Mt \text{ for some } M>0\}.$$

One can verify that C_τ^1 is a Banach space when equipped with the norm $|\cdot|_\tau$ defined by

$$|x|_\tau\equiv\sup_{0<t\leqq\tau}\frac{|\eta'(t)|+|z'(t)|}{t}.$$

Note the inequalities

$$|\eta'(t)|+|z'(t)|\leqq|x|_\tau t,\qquad |\eta(t)|+|z(t)|\leqq|x|_\tau\frac{t^2}{2}. \tag{1.4}$$

We define the linear operator L on C_τ^1 by

$$(Lx)(t)\equiv(\lambda_0\eta(\lambda_0^{-1}t),Bz(\lambda_0^{-1}t)),\qquad 0\leqq t\leqq\tau.$$

Since

$$(Lx)'(t)=(\eta'(\lambda_0^{-1}t),\lambda_0^{-1}Bz'(\lambda_0^{-1}t)),\qquad 0\leqq t\leqq\tau,$$

$$\begin{aligned}|(Lx)'(t)|&\leqq|\eta'(\lambda_0^{-1}t)|+\lambda_0^{-1}|Bz'(\lambda_0^{-1}t)|\\&\leqq|\eta'(\lambda_0^{-1}t)|+\lambda_0^{-1}\|B\||z'(\lambda_0^{-1}t)|\\&\leqq\max\{1,\lambda_0^{-1}\|B\|\}[|\eta'(\lambda_0^{-1}t)|+|z'(\lambda_0^{-1}t)|]\\&\leqq\max\{1,\lambda_0^{-1}\|B\|\}|x|_\tau\lambda_0^{-1}t\\&\leqq\max\{\lambda_0^{-1},\lambda_0^{-2}\|B\|\}|x|_\tau t,\end{aligned}$$

it follows that $Lx\in C_\tau^1$ and

$$\|L\|\leqq\max\{\lambda_0^{-1},\lambda_0^{-2}\|B\|\}\equiv\rho<1$$

provided $\|B\|<\lambda_0^2$.

Define an operator R on C_τ^1 by

$$R(x)=R(\eta,z)(t)\equiv(r_1(\lambda_0^{-1}t+\eta(\lambda_0^{-1}t),z(\lambda_0^{-1}t)),r_2(\lambda_0^{-1}t+\eta(\lambda_0^{-1}t),z(\lambda_0^{-1}t))).$$

Since we want to make use of the estimates (1.1) and (1.2) we define R only for those $x\in\frac{1}{\tau}$ with $|x|_\tau\leqq m$ (m and τ will be adjusted further in what follows) so that

$$|\lambda_0^{-1}t+\eta(\lambda_0^{-1}t)|+|z(\lambda_0^{-1}t)|\leqq\leqq\left(\frac{\tau}{\lambda_0}\right)\left[1+\frac{m}{2}\left(\frac{\tau}{\lambda_0}\right)\right]\leqq\delta. \tag{1.5}$$

We assume τ is small enough that the estimate above holds. Then

$$\begin{aligned}
|R(\eta,z)'(t)| &= \left|Dr_1\left(\lambda_0^{-1}t+\eta,z\right)\left(\lambda_0^{-1}+\lambda_0^{-1}\eta',\lambda_0^{-1}z'\right)\right| \\
&\quad + \left|Dr_2\left(\lambda_0^{-1}t+\eta,z\right)\left(\lambda_0^{-1}+\lambda_0^{-1}\eta',\lambda_0^{-1}z'\right)\right| \\
&\leqq C\lambda_0^{-1}\left[\left|\lambda_0^{-1}t+\eta\left(\lambda_0^{-1}tt\right)\right|+\left|z\left(\lambda_0^{-1}t\right)\right|\right]\left[\left|1+\eta'\left(\lambda_0^{-1}t\right)\right|+\left|z'\left(\lambda_0^{-1}t\right)\right|\right] \\
&\leqq C\lambda_0^{-1}\left[\lambda_0^{-1}t+m\frac{\lambda_0^{-2}t^2}{2}\right]\left[1+m\lambda_0^{-1}t\right] \\
&\leqq C\lambda_0^{-2}\left[1+\frac{m\lambda_0^{-1}\tau}{2}\right]\left[1+m\lambda_0^{-1}\tau\right]t.
\end{aligned}$$

It follows that $R(\eta,z)\in C_\tau^1$ and that $|R(\eta,z)|_\tau \leqq C\lambda_0^{-2}[1+m\lambda_0^{-1}\tau]^2, |(\eta,z)|_\tau \leqq m$.

Equation (1.3) is equivalent to $x=F(x)\equiv Lx+R(x)$, $x=(\eta,z)\in C_\tau^1$. We will show that F is a contractive self-map of $\overline{B_m(0)}\subseteq C_\tau^1$ for some m and τ. For fixed $m>0$ and τ sufficiently small that (1.5) holds we have for $|x|_\tau \leqq m$, $|F(x)|_\tau \leqq \rho m + C\lambda_0^{-2}[1+m\lambda_0^{-1}\tau]^2$. In order that F be a self-map of $\overline{B_m(0)}$ we choose

$$m=\frac{2C\lambda_0^{-2}}{1-\rho}$$

and $\tau<\min\{(\sqrt{2}-1)(1-\rho)/2C\lambda_0^3, \lambda_0(\sqrt{1+2m\delta}-1)/m\}$. The second term inside the min bracket has been chosen so that (1.5) holds while the first term insures that, together with our choice of m, F is a self map of $\overline{B_m(0)}$. We will need to choose τ possibly smaller to insure that F is a contraction.

Let $x_i=(\eta_i,z_i)\in\overline{B_m(0)}$, $i=1,2$. Then, using the estimate (1.2)

$$\begin{aligned}
&|R(x_1)'(t)-R(x_2)'(t)| \\
&= \sum_{i=1}^{2}\Big|Dr_i\left(\lambda_0^{-1}t+\eta_1\left(\lambda_0^{-1}t\right),z_1\left(\lambda_0^{-1}t\right)\right)\left(\lambda_0^{-1}+\lambda_0^{-1}\eta_1'\left(\lambda_0^{-1}\right),\lambda_0^{-1}z_1'\left(\lambda_0^{-1}t\right)\right) \\
&\qquad - Dr_i\left(\lambda_0^{-1}t+\eta_2\left(\lambda_0^{-1}t\right),z_2\left(\lambda_0^{-1}t\right)\right)\left(\lambda_0^{-1}+\lambda_0^{-1}\eta_2'\left(\lambda_0^{-1}t\right),\lambda_0^{-1}z_2'\left(\lambda_0^{-1}t\right)\right)\Big| \\
&\leqq C\Big[\lambda_0^{-1}\max_i\left|\left(\lambda_0^{-1}t+\eta_i\left(\lambda_0^{-1}t\right),z_i\left(\lambda_0^{-1}t\right)\right)\right|\left|x_i'\left(\lambda_0^{-1}t\right)-x_2'\left(\lambda_0^{-1}t\right)\right| \\
&\qquad + \max_i\left|\left(\lambda_0^{-1}+\lambda_0^{-1}\eta_i'\left(\lambda_0^{-1}t\right),\lambda_0^{-1}z_i'\left(\lambda_0^{-1}t\right)\right)\right|\left|x_1\left(\lambda_0^{-1}t\right)-x_2\left(\lambda_0^{-1}t\right)\right|\Big] \\
&\leqq C\left[\left(\lambda_0^{-1}t+m\frac{\lambda_0^{-2}t^2}{2}\right)\left(\lambda_0^{-2}t|x_1-x_2|_\tau\right)+\left(\lambda_0^{-1}+\lambda_0^{-2}tm\right)\frac{\lambda_0^{-2}t}{2}|x_1-x_2|_\tau\right] \\
&\leqq C\lambda_0^{-3}\tau\left[\left(1+\frac{m\lambda_0^{-1}\tau}{2}\right)+\frac{1}{2}\left(1+m\lambda_0^{-1}m\right)\right]|x_1-x_2|_\tau t \\
&\leqq C\lambda_0^{-3}\tau\left[\frac{3}{2}+m\lambda_0^{-1}\tau\right]|x_1-x_2|_\tau t.
\end{aligned}$$

It follows that

$$|R(x_1)-R(x_2)|_\tau \leqq C\lambda_0^{-3}\tau\left[\frac{3}{2}+m\lambda_0^{-1}\tau\right]|x_1-x_2|_\tau.$$

Hence

$$|F(x_1)-F(x_2)|_\tau \leqq \left[\rho+C\lambda_0^{-3}\tau\left[\frac{3}{2}+m\lambda_0^{-1}\tau\right]\right]|x_1-x_2|_\tau.$$

Since $\rho<1$, we can insure that the Lipschitz constant for F is smaller than one by further restricting τ if necessary. The contraction mapping theorem can now be applied completing the proof of the lemma.

Before continuing the proof of Theorem 1.1 we prove two results, Propositions 1.3 and 1.4 below, the proofs of which are very similar to the proof of Lemma 1.2. These results indicate several possible variations of Theorem 1.1. The first result, Proposition 1.3 below, shows that the smoothness assumptions on T in Lemma 1.2 can be relaxed with a corresponding loss of smoothness of y. The proof of the following lemma is so similar to the proof of Lemma 1.2 that only a sketch is given.

PROPOSITION 1.3. *Let T: $\Omega\to X$ as in Theorem* 1.1 *but assume only that T is continuous on Ω, $T(x_0)=x_0$ and $DT(x_0)=A$ exists. Let the spectral assumptions of Theorem* 1.1 *hold for A. Define $r(x)$ for x near x_0 by*

$$T(x)=x_0+A(x-x_0)r(x)$$

and suppose there exists $M,p,\delta>0$ such that

$$|r(x_1)-r(x_2)|\leqq C\left[\max_{i=1,2}\{|x_i-x_0|\}\right]^p|x_1-x_2| \tag{1.6}$$

for $|x_0-x_i|\leqq\delta$, $i=1,2$. Then there exists $\tau>0$ and a continuous function y: $[0,\tau]\to\Omega$ satisfying $y(t)=x_0+te_0+O(t^{1+p})$ as $t\to 0$, (ii) *and* (iii) *of Theorem* 1.1 *(except $x_n'\to y'$).*

Proof. The proof is very similar to that of Lemma 1.2. The estimates (1.1) and (1.2) of Lemma 1.2 are replaced by the single estimate

$$\sum_{i=1}^{2}|r_i(u_1,z_1)-r_i(u_2,z_2)|\leqq C\left[\max_{i=1,2}\{|u_i|+|z_i|\}\right]^p\cdot\left[|u_1-u_2|+|z_1-z_2|\right] \tag{1.7}$$

if $|u_i|+|z_i|\leqq\delta$, $i=1,2$, which follows from (1.6). The appropriate function space is

$$C_p=\{(\eta(t),z(t))\in C([0,\tau],X)\colon |\eta(t)|+|z(t)|\leqq Mt^{1+p}\text{ for some }M\geqq 0\}$$

with p as in (1.6). The norm on C_ρ is given by

$$|(\eta,z)|_p=\sup_{0<t\leqq\tau}\frac{|\eta(t)|+|z(t)|}{t^{1+p}}.$$

It is clear that $(C_\rho,\|_\rho)$ is a Banach space. Define L: $C_\rho\to C_\rho$ as in Lemma 1.2 and observe that $\|L\|\leqq\rho\equiv\max\{\lambda_0^{-p},\lambda_0^{-(1+p)}\|B\|\}$. Since $\|B\|<\lambda_0^{1+p}$ (this may require a different norm on Z than in Lemma 1.2), $\rho<1$. If $m>0$ is given, choose $\tau>0$ so that $\tau\lambda_0^{-1}[1+m(\tau\lambda_0^{-1})^p]\leqq\delta$. Then given $(\eta,z)\in\overline{B_m(0)}$, the estimate (1.7) can be used to show that $R(\eta,z)$, defined as in Lemma 1.2, satisfies

$$|R(\eta,z)|_p\leqq C\lambda_0^{-(1+p)}\left[1+m(\tau/\lambda_0)^p\right]^{1+p}.$$

Hence $F \equiv L+R$ is a self-map of $\overline{B_m(0)}$ provided $\rho m + C\lambda_0^{-(1+p)}[1+m(\tau/\lambda_0)^p]^{1+p} \leqq m$. As in Lemma 1.2, since $\rho<1$ and we are free to choose τ small, this inequality can be satisfied for suitable m and τ (for example, $m=2C^{-(1+p)}/1-\rho$). Now use (1.7) to calculate the contraction constant, $\rho + C\lambda_0^{-(1+p)}(\tau/\lambda_0)^p[1+-m(\tau/\lambda_0)^p]^p$, for F. For τ sufficiently small, this value is less than one.

As our final variant of Lemma 1.2, we consider the case that a complex conjugate pair of eigenvalues, $\alpha \pm i\beta$, of A are the dominant eigenvalues. This result would have a cleaner form if T were assumed to be a holomorphic map of a complex Banach space (in which case y below would be holomorphic) but in view of possible applications we continue to assume X is a real Banach space.

PROPOSITION 1.4. *Assume T: $\Omega \to X$, Ω an open subset of the Banach space X, is C^2 in a neighborhood of a fixed point x_0 of T. Let $A=DT(x_0)$ and assume*

(a) *$\lambda_0=\alpha+i\beta$ is a simple pole of the resolvent $R(\lambda,A^c)=(\lambda I-A^c)^{-1}$ and $\dim N(A^c-\lambda_0 I)<\infty$. A^c is the complexification of A.*

(b) *The spectral radius of A is $|\lambda_0|$ and $|\lambda_0|>1$.*

Then there exists $\tau>0$ and a C^1 function y: $B_\tau(0)\subseteqq R^2 \to \Omega$ such that

(i) *$y(0)=x^0$, $Dy(0)$ has rank two.*

(ii) *$y(v)=T(y(S^{-1}v))$, $v\in B_\tau(0)$ where $S=\begin{pmatrix}\alpha & -\beta\\ \beta & \alpha\end{pmatrix}$.*

Proof. By using the functional calculus of operators [2] as in the proof of Lemma 1.2, one can regard X as the direct sum

$$X=R^2+Z, \qquad x=(v,z),\ v\in R^2,\ z\in Z$$

where the decomposition reduces A: $A=S+B$, B: $Z\to Z$. Since the spectral radius of B does not exceed $|\lambda_0|$ we may choose an equivalent norm on X so that $|(v,z)|=|v|+|z|$ where $|v|$ is the Euclidean norm and $|z|$ is a norm on Z such that the operator norm $\|B\|<|\lambda_0|^2$. Note that $\|S\|=|\lambda_0|$ and $\|S^{-1}\|=|\lambda_0|^{-1}$.

Assume $x_0=0$ and write

$$T(x)=T(v,z)=(T_1(v,z),T_2(v,z))$$

where

$$T_1\colon R^2\times Z\to R^2 \quad \text{and} \quad T_2\colon R^2\times Z\to Z$$

satisfy

$$T_1(v,z)=Sv+r_1(v,z), \qquad T_2(v,z)=Bv+r_2(v,z).$$

Our assumptions on T insure that there exists δ, $C>0$ such that $|(v_i,z_i)|\leqq\delta$, $i=1,2$, implies

$$\sum_{i=1}^{2}\|Dr_i(v_1,z_1)-Dr_i(v_2,z_2)\|\leqq C|(v_1,z_1)-(v_2,z_2)|.$$

We seek y: $B_\tau(0)\subseteqq R^2\to X$ in the form $y(v)=(v+\eta(v),z(v))$ with $|(\eta(v),z(v))|=O(|v|^2)$ and such that (ii) holds. This amounts to

$$\eta(v)=S\eta(S^{-1}v)+r_1(S^{-1}v+\eta(S^{-1}v),z(S^{-1}v)),$$
$$z(v)=Bz(S^{-1}v)+r_2(S^{-1}v+\eta(S^{-1}v),z(S^{-1}v)).$$

We solve these equations as in Lemma 1.2 by the contraction mapping theorem in the function

$$C_\tau^1 = \{(\eta,z)\colon \overline{B_\tau(0)} \subseteqq R^2 \to R^2 \times Z\colon (\eta,z)$$
$$\text{is } C^1 \text{ in } \overline{B_\tau(0,0)},\ (\eta(0),z(0))=(0,0),$$
$$\text{and for some } M \geqq 0\ \|D\eta(v)\|+\|Dz(v)\| \leqq M|v|,\ v\in B_\tau(0)\}.$$

Note that $D\eta(v)\colon R^2 \to R^2$ and $Dz(v)\colon R^2 \to Z$.

A Banach space norm for C_τ^1 is

$$|(\eta,z)|_\tau = \sup_{0<|v|\leqq\tau} \frac{\|D\eta(v)\|+\|Dz(v)\|}{|v|}.$$

Observe that the inequality

$$|\eta(v)|+|z(v)| \leqq \frac{1}{2}|(\eta,z)|_\tau |v|^2$$

holds for $(\eta,z)\in C_\tau^1$. We now proceed exactly as in Lemma 1.2 (indeed, all the estimates are identical except that $|\lambda_0|$ replaces λ_0). For instance $L\colon C_\tau^1 \to C_\tau^1$ is defined by $(\bar\eta,\bar z)=L(\eta,z)=(S\eta(S^{-1}v), Bz(S^{-1}v))$. Since

$$\begin{aligned}
\|D\bar\eta(v)\|+\|D\bar z(v)\| &= \|D\eta(S^{-1}v)\|+\|BDz(S^{-1}v)S^{-1}\| \\
&\leqq \|D\eta(S^{-1}v)\|+\|B\|\|Dz(S^{-1}v)S^{-1}\| \\
&\leqq \|D\eta(S^{-1}v)\|+\|B\||\lambda_0|^{-1}\|Dz(S^{-1}v)\| \\
&\leqq \max\{1,\|B\||\lambda_0|^{-1}\}|(\eta,z)|_\tau |S^{-1}v| \\
&\leqq \max\{|\lambda_0|^{-1},\|B\||\lambda_0|^{-2}\}|(\eta,z)|_\tau |v|,
\end{aligned}$$

it follows that $\|L\| \leqq \max\{|\lambda_0|^{-1}, \|B\||\lambda_0\|^{-2}\} < 1$. We leave it to the reader to check that the other estimates hold verbatim as in Lemma 1.2.

We now complete the proof of Theorem 1.1.

Proof of Theorem 1.1. The first task is to extend our function $y\colon [0,\tau]\to\Omega$ of Proposition 1.2 as far as possible in such a way that (ii) holds. Extend y as follows if $0\leqq t\leqq\lambda_0\tau$ let $y_1(t)\equiv T(y(\lambda_0^{-1}t))$; since $0\leqq\lambda_0^{-1}t\leqq\tau$, $y(\lambda_0^{-1}t)$ is well-defined and belongs to Ω, so $y_1(t)$ is well-defined and C^1 since y and T are. Also, y_1 agrees with y on $[0,\tau]$ so $y_1(t)=T(y_1(\lambda_0^{-1}t))$ for $0\leqq t\leqq\lambda_0\tau$. However, it may happen that $y_1(t)\notin\Omega$ for all $t\in[0,\lambda_0\tau]$. Suppose that this is the case and let $\lambda_0^{-1}t_0\in(\tau,\lambda_0\tau]$ be maximal with the property that $y_1(t)\in\Omega$ for $[0,\lambda_0^{-1}t_0)$. If $0\leqq t<t_0$ define $y_2(t)=T(y_1(\lambda_0^{-1}t))$. It is again easy to see that y_2 is well-defined, C^1, and agrees with y_1 on $0\leqq t\leqq\lambda_0\tau$. Hence $y_0(t)=T(y_2(\lambda_0^{-1}t))$ for $0\leqq t<t_0$. In general, this is as far as we can extend y. If, however, $y_1(t)\in\Omega$ for $0\leqq t\leqq\lambda_0\tau$ then y_1 can be extended exactly as before to $[0,\lambda_0^2\tau]$. This completes the proof of (ii) and (iv).

In order to see that (iii) holds for $0<t<t_0$ (recall we have proved (iii) on $[0,\tau]$ in Lemma 1.2), fix $t<t_0$ and let n be such that $\lambda_0^{-n}t\leqq\tau$. We have

$$\begin{aligned}
y(t)&=T\big(y(\lambda_0^{-1}t)\big)=T\big(T\big(y(\lambda_0^{-2}t)\big)\big)=\cdots\\
&=T^n\big(y(\lambda_0^{-n}t)\big)\\
&=T^n\Big(\lim_{p\to\infty}T^p\big(\lambda_0^{-(n+p)}te_0\big)\Big)\\
&=\lim_{p\to\infty}T^{n+p}\big(\lambda_0^{-(n+p)}te_0\big)\\
&=\lim_{p\to\infty}x_{n+p}(t).
\end{aligned}$$

Since T is uniformly continuous on bounded sets, $x_n\to y$ uniformly on compact sets.

From above we have

$$\begin{aligned}
y'(t)&=D(T^n)\big(y(\lambda_0^{-n}t)\big)\lambda_0^{-n}y'(\lambda_0^{-n}t)\\
&=D(T^n)\lim_{p\to\infty}x_p(\lambda_0^{-n}t)\Big)\lambda_0^{-n}\lim_{p\to\infty}x_p'(\lambda_0^{-n}t)\\
&=\lim_{p\to\infty}D(T^n)\big(x_p(\lambda_0^{-n}t)\big)\lambda_0^{-n}x_p'(\lambda_0^{-n}t)\\
&=\lim_{p\to\infty}x_{n+p}'(t).
\end{aligned}$$

This completes the proof of (iii).

Finally, if T is a diffeomorphism and there exists $s\neq t$ such that $y(t)=y(s)$ then $T(y(\lambda_0^{-1}t))=T(y(\lambda_0^{-1}s))$ and $y(\lambda_0^{-1}t)=y(\lambda_0^{-1}s)$. After n applications of this reasoning $y(\lambda_0^{-n}t)=y(\lambda_0^{-n}s)$ which contradicts (for n sufficiently large) the fact that y is one-to-one on $[0,\varepsilon]$ for $\varepsilon>0$ sufficiently small (recall y is C^1 and $y'(0)=e\neq 0$). Since $y'(t)\neq 0$ for small t there is a maximal $t_1\leqq t_0$ such that $y'(t)=0$ on $[0,t_1)$. If $t_1<t_0$ then $y'(t_1)=0$. But then

$$0=y'(t_1)=\lambda_0 DT\big(y(\lambda_0^{-1}t_1)\big)y'(\lambda_0^{-1}t_1),$$

yet $y'(\lambda_0^{-1}t_1)\neq 0$ and $DT(y(\lambda_0^{-1}t_1))$ is nonsingular. This contradiction proves $t_1=t_0$ and $y'(t)\neq 0$ on $[0,t_0)$.

2. Some results for monotone mappings. We begin by recalling some notation and results in the theory of partially ordered spaces. Recall that a cone K in a Banach space X is a closed subset of X with the properties (i) $K+K\subset K$, (ii) $R_+\cdot K\subset K$, and (iii) $K\cap(-K)=\{0\}$. K induces a partial ordering on X via $x\leqq y$ if and only if $y-x\in K$. If $x\leqq y$ are two points in X we write $[x,y]$ for the set $\{z\in X\colon x\leqq z\leqq y\}$. A map $F\colon X\to Y$ between two Banach spaces containing cones K and C, respectively, is nondecreasing provided $F(x)\leqq_c F(y)$ in Y whenever $x\leqq_k y$ in X. A cone in a Banach space X is said to be normal if there is an equivalent nondecreasing norm on X. It is easily verified that the Fréchet derivative of a nondecreasing map $T\colon X\to X$ at a point x, $A=DT(x)$, is a so-called positive operator, that is, $A(K)\subset K$. A positive operator A is said to be strongly positive if $A^m(K\subset\{0\})\subset\operatorname{int}K$, the interior of K, for some positive integer $m\geqq 1$ (of course, this makes sense only if K has nonempty interior). In case $X=R^n$ and $K=R_+^n=\{x\colon x_i\geqq 0, 1\leqq i\leqq n\}$, a nonnegative matrix A is strongly positive if and only if all entries of A^m are positive for some integer $m\geqq 1$. In this case, the Perron–Frobenius

Theory [6] [9] implies that the spectral radius is a simple eigenvalue, greater in magnitude than any other eigenvalue, and there is a corresponding eigenvector which belongs to the interior of R^n_+. A slightly weaker assumption on $A \geqq 0$ for which the above results hold is that A be a primitive irreducible matrix (see [9]), however these ideas do not generalize nicely to infinite dimensions. For infinite-dimensional spaces X and cone $K \subset X$ with nonempty interior, strongly positive compact operators have the properties mentioned above for strongly positive matrices except possibly that there may be other parts of the spectrum in addition to the spectral radius on the circle of radius equal to the spectral radius in the complex plane [1, Thm. 3.2].

Recall that T: $X \to X$, X a Banach space, is said to be completely continuous if T maps bounded sets into precompact sets. It is well known that the Fréchet derivative of a completely continuous map is a compact linear operator.

We can now state the main results of this section.

THEOREM 2.1. *Let K be a normal cone with nonempty interior in a Banach space X. Let T: $X \to X$ be a completely continuous, nondecreasing, C^1-map, which is C^2 in a neighborhood of 0 and $T(0)=0$. Let $A=DT(0)$ be strongly positive and have spectral radius λ_0, $\lambda_0 > 1$. Then there exists a unique C^1, nondecreasing function y: $[0,\infty] \to \operatorname{int} K$ satisfying $y(t)=te_0+o(t)$ as $t \to 0$ where $Ae_0=\lambda_0$, $e_0 \in \operatorname{int} K$ and*

$$y(t)=T\big(y\big(\lambda_0^{-1}t\big)\big),\ t \geqq 0.$$

Moreover, either $|y(t)| \to +\infty$ as $t \to \infty$ or $\lim_{t\to\infty} y(t)=y_\infty \in \operatorname{int} K$ exists and $Ty_\infty = y_\infty$. In the latter case $[0,y_\infty]$ is invariant under T and $T^n_x \to y_\infty$ for every $x \in [0,y_\infty]-\{0\}$. In the former case, $|T^n x| \to +\infty$ for all $x \in K-\{0\}$.

An important extension of Theorem 2.1 can be made if $X=R^n$ and $K=R^n_+$. In this case, the complete continuity assumption in Theorem 2.1 is redundant since T is continuous. We let $\rho(A)$ stand for the spectral radius of a linear operator A.

THEOREM 2.2. *Let $X=R^n$ and $K=R^n_+$ and T: $R^n \to R^n$ satisfy the hypotheses of Theorem 2.1 and let $DT(x)$ be invertible for $x \in R^n_+$. Assume $\lim_{t\to\infty} y(t)=y_\infty$ exists and assume $A_\infty \equiv DT(y_\infty)$ is strongly positive. Then $\rho(A_\infty) \leqq 1$ and $\lim_{t\to\infty}(y'(t)/|y'(t)|)=e$ where $e \in \operatorname{int} R^n_+$ is the unique, unit norm positive eigenvector corresponding to $\rho(A_\infty)$.*

Before proceeding to the proof of the theorems we make the following remarks. First the smoothness assumptions on T in Theorem 2.1 can be relaxed as observed in Remark 6 following Theorem 1.1. Secondly, the choice of the origin as a fixed point of T in both theorems is merely for convenience. Finally, (iii) of Theorem 1.1 holds for y in both results.

For diffeomorphisms T: $R^n \to R^n$, nondecreasing with respect to the usual ordering and having the property that $DT(x)$ is strongly positive for each x, Theorems 2.1 and 2.2 allow the determination of possible "phase portraits" for the dynamics, $x_{n+1}=T(x_n)$ in case of the existence of an unstable fixed point x_0 of $T(\rho(DT(x_0))>1)$. In Fig. 1 we indicate the four possible portraits. If $e_0 \in \operatorname{int} R^n_+$ satisfies $DT(x_0)e_0=\rho(DT(x_0)e_0$, we write $y_+(t)$ for the function the existence of which is asserted in Theorem 2.1 corresponding to the eigenvector $e_0(y_+(t)=x_0+te_0+O(t^2))$ and write $y_-(t)$ for the function asserted to exist in Theorem 2.1 corresponding to $-e_0(y_-(t)=x_0-te_0+O(t^2))$. by Theorem 2.1, $y_\pm(t) \in x_0 \pm \operatorname{int} R^n_+$ for each $t>0$. The four possibilities indicated in Fig. 1 correspond to the four possible limits $\lim y_+(t)=y_\infty$ or ∞, $\lim y_-(t)=y_{-\infty}$ or ∞. It should be observed that, in case $\lim_{t\to\infty} y_+(t)=y_\infty$ exists, the fixed point y_∞ may be the limit $\lim_{t\to\infty} y_-(t)$, of a function $y_-(t)$ corresponding to another unstable fixed point of T belonging to $y_\infty + \operatorname{int} R^n_+$. A similar observation holds if $\lim_{t\to\infty} y_-(t)=y_{-\infty}$

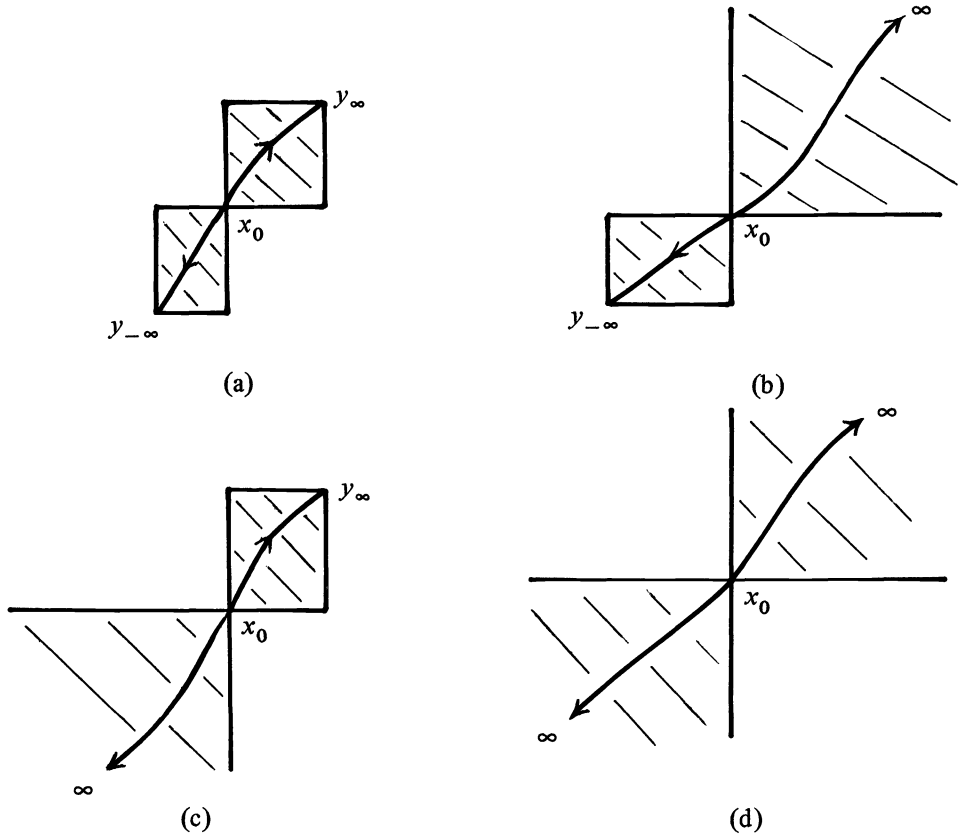


FIG. 1. *Qualitative dynamics of the iterates of* T: $R^2 \to R^2$; x_0 *is an unstable fixed point, shaded regions are included in the domains of attraction of a fixed point* $y_{-\infty}$, y_∞ *or the point at infinity,* ∞. *The solid curves are* y_+ *and* y_-.

exists. The point here is that (a), (b), and (c) of Fig. 1 may only be segments of a longer chain of such segments. A nice example of these chains occurs in the paper of Selgrade [8, see Figs. 1 and 2] where T is the time-one map for autonomous ordinary differential equation.

Finally, note that if x_0 is a a hyperbolic fixed point of T and $W^s(x_0) = \{x \in R^n: T^m(x) \to x_0 \text{ as } m \to \infty\}$ is the stable manifold then $W^s(x_0) \cap [(x_0 + R^n_+) \cup (x_0 - R^n_+)] = \{x_0\}$. This follows immediately from Theorem 2.1.

Proof of Theorem 2.1. As noted above, the spectral assumptions on A of Theorem 1.1 are satisfied by virtue of [1, Thm. 3.2] (notice that since λ_0 is a simple eigenvalue of A, λ_0 is a simple pole of $(A - \lambda I)^{-1}$, see Remark 1). By Theorem 1.1, there exists a unique function y: $[0, \infty) \to X$ satisfying the functional equation $y(t) = T(y(\lambda_0^{-1}t))t \geqq 0$. That $y(t) \in K$ follows since $T^n(\lambda_0^{-n}te_0) \in K$ for $n = 1, 2, \cdots, t \geqq 0$, and K is closed. Indeed $T^n(\lambda_0^{-n}te_0)$ is a monotone nondecreasing function of t for each $n = 1, 2, \cdots$ and thus so is $y(t)$. It is easy to see that $y(t) \in \operatorname{int} K$ for small t and since $y(t)$ is nondecreasing, it lies in $\operatorname{int} K$ for all t. If $\{y(t): 0 \leqq t < \infty\}$ is bounded then it is precompact since T is completely continuous. Let $t_n \to +\infty$ and $s_n \to +\infty$ and suppose $\lim y(t_n) = y_1$ and $\lim y(s_n) = y_2$ exist. Since $y(t)$ is monotone nondecreasing we have $y_1 \geqq y(s_n)$ for $n = 1, 2, \cdots$ so $y_1 \geqq y_2$ and similarly $y_2 \geqq y_1$. It follows that $y_1 = y_2$. Since $\{y(t): 0 < t < \infty\}$ is precompact, from every sequence $t_n \to \infty$ there is a subsequence

which converges. The above observation shows that the limit of any such subsequence is the same point. Hence $\lim_{t\to\infty} y(t)=y_\infty$ exists and is a fixed point by continuity of T. Since $y_\infty \geqq y(t)\in \operatorname{int} K$, $y_\infty \in \operatorname{int} K$. It is clear that $[0,y_\infty]$ is invariant under T since 0 and y_∞ are fixed by T and T is nondecreasing. If $x\in[0,y_\infty]\cap \operatorname{int} K$ then $y(t_0)\leqq x$ for small t_0 so $y(\lambda_0^n t_0)\leqq T^n x\leqq y_\infty$ after n applications of T to the previous inequality. Since $[0,y_\infty]$ is bounded (here we use normality of K) and T is completely continuous it follows that $T^n x\to y_\infty$. Now, if $x\in[0,y_\infty]-\{0\}$ then $T^m x\in[0,y_\infty]\cap \operatorname{int} K$ where m is such that $A^m(K-\{0\})\subseteq \operatorname{int} K$.

In case $\lim|y(t)|=+\infty$ and $x\in \operatorname{int} K$ then, again, $y(t_0)\leqq x$ for small t_0 so $y(\lambda_0^n t_0)\leqq T^n x$ as above. Since we may assume $|\cdot|$ to be nondecreasing, $|y(\lambda_0^n t_0)|\leqq|T^n x|$ for $n=1,2,\cdots$. It follows that $\lim|T^n x|=+\infty$.

Proof of Theorem 2.2. If $\rho(A_\infty)>1$ then since A_∞ is strongly positive we could use Theorem 1.1 to obtain a function y_-: $[0,\infty)\to[0,y_\infty]\cap \operatorname{int} K$ satisfying $y_-(0)=y_\infty$, $A_\infty y'_-(0)=\rho y'_-(0)$, $y_-(t)=T(y_-(\rho^{-1}t))$, $\rho=\rho(A_\infty)$. Clearly this contradicts that $[0,y_\infty]\cap \operatorname{int} K$ lies in the basin of attraction of y_∞.

Observe that $y(t)$ satisfies the functional differential equation

$$y'(t)=\lambda_0^{-1}DT\big(y(\lambda_0^{-1}t)\big)y'(\lambda_0^{-1}t) \tag{2.1}$$

and since $DT(x)$ is invertible it follows (Theorem 1.1) that $y'(t)\neq 0$ for all t. We have

$$\frac{y'(t)}{|y'(\lambda_0^{-1}t)|}=\lambda_0^{-1}A_\infty\frac{y'(\lambda_0^{-1}t)}{|y'(\lambda_0^{-1}t)|}+\lambda_0^{-1}\big[DT\big(y(\lambda_0^{-1}t)-A_\infty\big]\frac{y'(\lambda_0^{-1}t)}{|y'(\lambda_0^{-1}t)|} \tag{2.2}$$

where the second term on the right tends to zero at $t\to\infty$.

Define

$$Q=\left\{v\colon v=\lim_{n\to\infty}\frac{y'(t_n)}{|y'(t_n)|}\text{ for some sequence } t_n\to\infty\right\}.$$

To show that $\lim_{t\to\infty}(y'(t)/|y'(t)|)=e$ it suffices to show that $Q=\{e\}$. It is easy to check that Q is a nonempty compact set of unit vectors belonging to R_+^n, the latter since $y(t)$ is nondecreasing. If $v\in Q$, $v=\lim_{n\to\infty}(y'(t_n)j/|y'(t_n)|)$, it follows from putting $\lambda_0^{-1}t=t_n$ in (2.2) and taking limits that $\lim_{n\to\infty}(y'(\lambda_0 t_0)/|y'(t_n|)=\lambda_0^{-1}A_\infty v$. Hence

$$\lim_{n\to\infty}\frac{y'(\lambda_0 t_n)}{|y'(\lambda_0 t_n)|}=\frac{A_\infty v}{|A_\infty v|}\in Q.$$

We may define a continuous map f: $Q\to Q$ by $f(v)=A_\infty v/|A_\infty v|$ and observe that f is one to one since A_∞ is invertible. f is a homeomorphism of Q onto Q. For, if $v=\lim_{n\to\infty}(y'(t_n)/|y'(t_n)|)\in Q$, we may put $t=t_n$ in (2.1), divide both sides by $|y'(t_n)|$ and take limits to show that $\lim_{n\to\infty}(y'(\lambda^{-1}t_n)/|y'(t_n)|)=\lambda_0 A_\infty^{-1}v$. It follows that

$$\lim_{n\to\infty}\frac{y'(\lambda_0^{-1}t_n)}{|y'(\lambda_0^{-1}t_n)|}=\frac{A_\infty^{-1}v}{|A_\infty^{-1}v|}=v_1\in Q$$

and $f(v_1)=v$.

We now proceed to show that $f^n(v)\to e$ as $n\to\infty$ (f^n denoting n-fold composition as usual) uniformly for $v\in Q$. This is compatible with f^n: $Q\to Q$ being a homeo-

morphism for $n=1,2,\cdots$, only if $Q=\{e\}$. The principal tool in showing that $f^n(v)\to e$ as $n\to\infty$ is the fact that

$$f^n(v)=\frac{A_\infty^n v}{|A_\infty^n v|}=\frac{\rho^n}{|A_\infty^n v|}\cdot\frac{A_\infty^n}{\rho^n}v$$

and the Frobenius theorem [6, Appendix, Thm. 2.3]; $\lim_{n\to\infty}(A_\infty^n/\rho^n)=P$ where P is the projection onto the linear span of e defined $Px=(h\cdot x)e$ where $h\cdot e=1$, $A_\infty^t h=\rho h$, $h\in \operatorname{int} R_+^n$ ($\rho=\rho(A_\infty)$). Since Q is a compact subset of $R_+^n S^{n-1}$, $h\cdot x$ is bounded below by a positive constant uniformly for $x\in Q$. It follows from the above observations that $f^n(v)\to Pv/|Pv|=e$ uniformly for $v\in Q$ completing the proof of Theorem 2.2.

REFERENCES

[1] H. AMANN, *Fixed point equations and nonlinear eigenvalue problems in ordered Banach spaces*, SIAM Rev., 18 (1976), pp. 620–709.

[2] N. DUNFORD AND J. I. SCHWARTZ, *Linear Operators Part* I, Interscience, New York, 1957.

[3] J. GUCKENHEIMER AND P. HOLMES, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Applied Math. Sciences, Vol. 42, Springer-Verlag, New York, 1983.

[4] J. HADAMARD, *Sur l'itération et les solutions asymptotiques des équations differentielles*, Bull. Soc. Math. France 29, 1901, pp. 224–228.

[5] P. HARTMAN, *Ordinary Differential Equations*, P. Hartman, Baltimore, MD, 1973.

[6] S. KARLIN AND H. M. TAYLOR, *A First Course in Stochastic Processes*, 2nd ed., Academic Press, New York, 1975.

[7] J. R. POUNDER AND T. D. ROGERS, *The geometry of chaos: dynamics of a nonlinear second order difference equation*, Bull. Math. Biol., 42 (1980), pp. 551–597.

[8] J. E. SELGRADE, *Asymptotic behavior of solutions to single loop positive feedback systems*, J. Differential Equations, 38 (1980), pp. 80–103.

[9] R. S. VARGA, *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1963.

[10] J. A. YORKE AND T. Y. LI, *Period three implies chaos*, Amer. Math. Monthly, 82 (1975), pp. 985–992.

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

[12] M. W. HIRSCH, *Systems of differential equations which are competitive or cooperative* I: *limit sets*, this Journal, 13 (1982), pp. 167–179.

[13] _____, *Systems of differential equations which are competitive or cooperative* II: *convergence almost everywhere*, this Journal, 16 (1985), pp. 423–439.

[14] _____, *Stability and convergence in strongly monotone flows*, preprint.

[15] _____, *Attractors for discrete-time monotone dynamical systems in strongly ordered spaces*, Proc. Special Year in Geometry 1983–84, Univ. Maryland, College Park, to appear.

[16] H. L. SMITH, *Periodic solutions of periodic competitive and cooperative systems*, this Journal, 17 (1986), to appear.

[17] _____, *Periodic competitive differential equations and the discrete dynamics of competitive maps*, J. Differential Equations, to appear.

[18] R. L. ADLER AND T. J. RIVLIN, *Ergodic and mixing properties of Chebyshev polynomials*, Proc. American Mathematical Society 15, 1964, pp. 794–796.

SIAM J. MATH. ANAL.
Vol. 17, No. 5, September 1986

003

BREAKDOWN OF STABILITY IN SINGULARLY PERTURBED AUTONOMOUS SYSTEMS II. ESTIMATES FOR THE SOLUTIONS AND APPLICATION*

K. NIPP†

Abstract. In Nipp, [this Journal, 17 (1986), pp. 512–532] error estimates were derived for the trajectories of a singularly perturbed autonomous system. The results are extensions of a classical result due to A. N. Tikhonov. In this paper we give the transfer of those estimates to the solutions of the autonomous system and an application to a problem in biomathematics.

Special cases of our results are treated by Lebovitz and Schaar in Stud. Appl. Math., 54 (1975), pp. 229–260, 56 (1977), pp. 1–50, respectively. In order to transfer their estimates to the solutions, however, they introduce an artificial condition excluding a whole class of problems containing, e.g., the van der Pol relaxation oscillator as well as the example given in this paper.

Key words. ordinary differential equations, singular perturbations, breakdown of stability, nerve impulse equations

AMS(MOS) subject classifications. Primary 34E, 34D

1. Formulation of the problem. We first repeat the precise formulation of the problem. For introduction and motivation the reader is referred to [1].

Consider the autonomous system

$$\begin{aligned} \dot{x} &= f(x,y)+\varepsilon f^1(x,y,\varepsilon), \\ \varepsilon\dot{y} &= g(x,y)+\varepsilon g^1(x,y,\varepsilon), \end{aligned} \tag{1}$$

where x and y are m- and n-vectors, respectively, and $\varepsilon\in[0,\varepsilon_0]$, $\varepsilon_0<1$. Moreover, let all functions be sufficiently smooth in the domains considered. The corresponding reduced system is

$$\dot{x}=f(x,y), \qquad 0=g(x,y). \tag{2}$$

It is a well-known result, due to A. N. Tikhonov, that "corresponding solutions" of the systems (1) and (2) are close to each other if a certain stability assumption is satisfied (cf. [1]). We are interested in the situation where a solution of the reduced system loses its stability, and we will state a local result valid in a neighborhood of the point $(x,y)=(0,0)$ where we assume that the stability breaks down. As seen in [1] two cases ("$s<0$", "$s>0$") have to be considered depending on whether the flow along the reduced trajectory approaches the point $(0,0)$ or leads away from it. We will now state the assumptions in the first case; those in the second case are completely analogous (compare [1]).

A1. $$f_1(0,0)>0, \qquad f_k(0,0)=0 \qquad (k=2,3,\cdots,m),$$
$$g(0,0)=0.$$

Hence, there exist a domain $U\subset\mathbb{R}^{m+n}$, containing the origin, and a positive constant ρ such that $f_1(x,y)>\rho$ for $(x,y)\in U$.

* Received by the editors September 3, 1984, and in final form July 29, 1985.
† Seminar für Angewandte Mathematik, ETH-Zurich, CH-8092 Zurich, Switzerland.

A2.

$$g_y(0,0)=\begin{pmatrix} 0 & 0 & \cdots & 0 \\ 0 & & & \\ \vdots & & \bar{A} & \\ 0 & & & \end{pmatrix}$$

where all the eigenvalues of the $(n-1)\times(n-1)$ matrix $\bar{A}$ have negative real parts.

Moreover, let $\bar{y}:=(y_2,\cdots,y_n)$. As seen in [1], we may assume without loss of generality that $\bar{g}(x,y_1,\bar{y})$ has a Taylor formula beginning in y_1 with a cubic term. A2 implies that there exists a unique solution $\bar{w}(x,y_1)$ of $\bar{g}(x,y_1,\bar{y})=0$ in a neighborhood $\hat{U}\subset U$ containing the origin, having continuous partial derivatives there and satisfying $\bar{w}(0,0)=0$. For the remaining equation

$$g_1(x,y_1,\bar{w}(x,y_1))=0$$

we require

A3. There exists a function $\phi_1(x)$, which is defined and continuous for $(x_1,\bar{x})\in J\times\bar{\Omega}\subset\hat{U}$, where $J:=(s_1,0]$ for some $-1<s_1<0$, $\bar{\Omega}$ some neighborhood of $\bar{x}=0$, has continuous partial derivatives with respect to $\bar{x}$ there, and is C^1 in $J'\times\bar{\Omega}$, $J':=(s_1,0)$, satisfying

$$g_1(x,\phi_1(x),\bar{w}(x,\phi_1(x)))=0, \qquad x\in J\times\bar{\Omega},$$
$$\phi_1(0)=0,$$

and

$$\phi_1(x)=a_1(-x_1)^{\alpha}+p(x), \qquad x\in J\times\bar{\Omega},$$
$$\frac{\partial\phi_1}{\partial x_1}(x)=a_{11}(-x_1)^{\alpha-1}+p_1(x), \qquad x\in J'\times\bar{\Omega},$$

where $\alpha>0$ and a_1, $a_{11}\neq 0$, $p(0)=0$, $p(x_1,0)=o((-x_1)^{\alpha})$ and $p_1(x_1,0)=o((-x_1)^{\alpha-1})$ as $x_1\to 0^-$.

If $\alpha\geqq 1$ we suppose that $\phi_1(x)\in C^1(J\times\bar{\Omega})$.

Now let $\bar{\phi}(x):=\bar{w}(x,\phi_1(x))$ and $\phi(x):=(\phi_1,\bar{\phi})$ and consider the initial value problem

$$\frac{d\bar{x}}{dx_1}=\frac{\bar{f}(x_1,\bar{x},\phi(x_1,\bar{x}))}{f_1(x_1,\bar{x},\phi(x_1,\bar{x}))}, \qquad \bar{x}(0)=0 \tag{3}$$

for $(x_1,\bar{x})\in J\times\bar{\Omega}$.

Moreover, we introduce the following initial conditions to the system (1):

$$\begin{aligned} x(0,\varepsilon)&=x^0(\varepsilon)=X^0+O(\varepsilon),\\ y(0,\varepsilon)&=y^0(\varepsilon)=Y^0+O(\varepsilon), \end{aligned} \tag{4}$$

where X^0, Y^0 independent of ε and $(x^0(\varepsilon),y^0(\varepsilon))\in\hat{U}$, $x^0(\varepsilon)\in J'\times\bar{\Omega}$ for all $\varepsilon\in[0,\varepsilon_0]$.

We also suppose the following:

A4. The solution $\bar{U}(x_1)$ of (3) exists for $x_1\in J$, and

$$\left|\bar{x}^0(\varepsilon)-\bar{U}\left(x_1^0(\varepsilon)\right)\right|<c_0\varepsilon,$$
$$\left|y^0(\varepsilon)-V\left(x_1^0(\varepsilon)\right)\right|<c_0\varepsilon,$$

where $V(x_1):=\phi(x_1,\bar{U}(x_1))$.

There is one more condition, the essential stability condition that, together with A2, replaces the corresponding assumption in the Tikhonov case (see [1]).

A5. There are positive constants k and q such that

$$g_{1,y_1}\big(x_1, \bar{U}(x_1), V(x_1)\big) \leqq -k(-x_1)^q, \qquad x_1 \in J.$$

Let $X_1(t)$ be the solution of

$$\frac{dx_1}{dt} = f_1\big(x_1, \bar{U}(x_1), V(x_1)\big), \qquad x_1(0) = X_1^0.$$

It exists and is unique on an interval $[0, T]$, and increases there from X_1^0 to 0. Then

$$\big(X(t), Y(t)\big) := \big(X_1(t), \bar{U}(X_1(t)), V(X_1(t))\big)$$

is a solution of the reduced system (2) for $t \in [0, T]$ satisfying

$$\begin{aligned} &|X(0) - x^0(\varepsilon)| = O(\varepsilon), \\ &|Y(0) - y^0(\varepsilon)| = O(\varepsilon), \\ &X(T) = 0, \qquad Y(T) = 0. \end{aligned} \tag{5}$$

Remark. • A1–A5 imply the corresponding assumptions in [1].

• Assumption A4, which is equivalent to the condition (5) for a solution of the reduced system (2), is naturally satisfied if we consider a global problem whose reduced trajectory approaches the point $(x,y) = (0,0)$ and is stable as long as it is in a finite distance from this point.

• In order to verify A5 the asymptotic relations for $(\bar{U}(x_1), V(x_1))$ given in [1] may be used. □

In order to be able to formulate our result we need define the following nonnegative quantities:

$$\begin{aligned} \hat{\alpha} &:= \min(\alpha, 1), \\ \bar{\alpha} &:= \min(2\alpha, 1), \\ \beta &:= \begin{cases} \alpha & \text{if } \alpha < 1 \text{ and } g_1(x,y) \text{ has no linear term in } x_1 \\ 0 & \text{otherwise;} \end{cases} \end{aligned}$$

moreover, γ and ν which are defined by the estimates

$$\begin{aligned} &\big|g_{1,\bar{x}}\big(x_1, \bar{U}(x_1), V(x_1)\big)\big| \leqq C(-x_1)^{\gamma}, \\ &\big|g_{1,\bar{y}}\big(x_1, \bar{U}(x_1), V(x_1)\big)\big| \leqq C(-x_1)^{\nu}, \qquad x_1 \in J. \end{aligned}$$

In [1] we have shown that

$$\begin{aligned} &\gamma \begin{cases} \geqq \hat{\alpha} & \text{if } \alpha \geqq 1 \text{ or } g_1(x,y) \text{ has no linear term in } \bar{x}, \\ = 0 & \text{otherwise,} \end{cases} \\ &\nu \geqq \hat{\alpha}. \end{aligned}$$

2. The main result. Under the assumptions A1–A5 the following result holds.

THEOREM 1. *If q satisfies*

$$q < \min(1 + \gamma, \bar{\alpha} + \nu)$$

then for $|x^0(\varepsilon)|$, $|y^0(\varepsilon)|$ *sufficiently small there exist positive constants* C *and* $\varepsilon_1 \leqq \varepsilon_0$ *such that the solution* $(x(t,\varepsilon),y(t,\varepsilon))$ *of the initial value problem* (1), (4) *exists at least for* $t \in [0, T - C\varepsilon^{q}*]$, *where* $q^* := 1/(q_1+q-\beta)$, $q_1 := 1-\hat{\alpha}+q$, *and*

$$|x(t,\varepsilon)-X(t)| < \begin{cases} C\varepsilon \log(T-t)^{-1}, & q_1 = 1, \\ C\varepsilon(T-t)^{1-q_1}, & q_1 > 1, \end{cases}$$

$$|y_1(t,\varepsilon)-Y_1(t)| < C\varepsilon(T-t)^{-q_1},$$

$$|\bar{y}(t,\varepsilon)-\bar{Y}(t)| < \begin{cases} C\varepsilon \log(T-t)^{-1}, & q_1 = 1, \\ C\varepsilon(T-t)^{\bar{\alpha}-q_1}, & q_1 > 1 \end{cases}$$

for $t \in [0, T - C\varepsilon^{q}*]$, $\varepsilon \in [0, \varepsilon_1]$. □

Proof. We want to apply [1, Thm. 2]. Consider the initial value problem

$$\begin{aligned} &\frac{d\bar{x}}{dx_1} = F(x_1,\bar{x},y) + \varepsilon F^1(x_1,\bar{x},y,\varepsilon), \qquad \bar{x}\big(x_1^0(\varepsilon),\varepsilon\big) = \bar{x}^0(\varepsilon), \\ &\varepsilon\frac{dy}{dx_1} = G(x_1,\bar{x},y) + \varepsilon G^1(x_1,\bar{x},y,\varepsilon), \qquad y\big(x_1^0(\varepsilon),\varepsilon\big) = y^0(\varepsilon), \end{aligned} \tag{6}$$

where $F := \bar{f}/f_1$, $G := g/f_1$, F^1 and G^1 are defined for $(x,y) \in U$. Let $(\bar{u}(x_1,\varepsilon), v(x_1,\varepsilon))$ be the solution of (6). By means of [1, Thm. 2] it exists for $x_1 \in J^* := [x_1^0(\varepsilon), -c\varepsilon^{q}*]$ if $|x_1^0(\varepsilon)|$ is taken small enough. And let $\tilde{x}_1(t,\varepsilon)$ be the solution of

$$\frac{dx_1}{dt} = f_1\big(x_1,\bar{u}(x_1,\varepsilon),v(x_1,\varepsilon)\big) + \varepsilon f_1^1\big(x_1,\bar{u}(x_1,\varepsilon),v(x_1,\varepsilon),\varepsilon\big), \qquad x_1(0,\varepsilon) = x_1^0(\varepsilon).$$

Then the following identity holds

$$(x(t,\varepsilon),y(t,\varepsilon)) = \big(\tilde{x}_1(t,\varepsilon),\bar{u}\big(\tilde{x}_1(t,\varepsilon),\varepsilon\big),v\big(\tilde{x}_1(t,\varepsilon),\varepsilon\big)\big)$$

as long as $\tilde{x}_1(t,\varepsilon) \in J^*$. Moreover, assumption A1 implies that $x_1(t,\varepsilon)$ is an, in t, increasing function there. Since this is also true for the reduced solution $X_1(t)$ (cf. §1), we may consider the inverse functions which satisfy the following integral equations. To save writing we put $s := x_1$, $s^0 := x_1^0(\varepsilon)$, $S^0 := X_1^0$, $s^* := -c\varepsilon^{q}*$, and without loss of generality we suppose that $S^0 > s^0$ for all $\varepsilon \in [0, \varepsilon_0]$.

$$X_1^{-1}(s) = \int_{S^0}^{s} \frac{d\sigma}{f_1\big(\sigma,\bar{U}(\sigma),V(\sigma)\big)}, \qquad s \in [S^0, 0],$$

$$\begin{aligned} x_1^{-1}(s,\varepsilon) &= \int_{s^0}^{s} \frac{d\sigma}{f_1\big(\sigma,\bar{u}(\sigma,\varepsilon),v(\sigma,\varepsilon)\big) + \varepsilon f_1^1\big(\sigma,\bar{u}(\sigma,\varepsilon),v(\sigma,\varepsilon),\varepsilon\big)}, \qquad s \in [s^0, s^*] \\ &= \int_{s^0}^{s} \frac{d\sigma}{f_1\big(\sigma,\bar{U}(\sigma)+z(\sigma,\varepsilon),V(\sigma)+w(\sigma,\varepsilon)\big) + \varepsilon f_1^1(\cdots)}. \end{aligned}$$

We define $X_1^{-1}(s) \equiv 0$ for $s < S^0$. Hence (cf. A1), for ε small enough there are positive constants M_1 and M_2 such that

$$|x_1^{-1}(s,\varepsilon) - X_1^{-1}(s)| < M_1\varepsilon + M_2\int_{s^0}^{s} \big(|z(\sigma,\varepsilon)| + |w(\sigma,\varepsilon)|\big)\,d\sigma.$$

Applying [1, Thm. 2] and integrating yields

$$\left|x_1^{-1}(s,\varepsilon)-X_1^{-1}(s)\right|<M_1\varepsilon+M_3\varepsilon\psi(s)<\hat{M}\varepsilon\psi(s),\qquad s\in[s^0,s^*]$$

where

$$\psi(s):=\begin{cases}\log(-s)^{-1}, & q_1=1,\\ (-s)^{1-q_1}, & q_1>1.\end{cases}$$

This situation is sketched in Fig. 1.

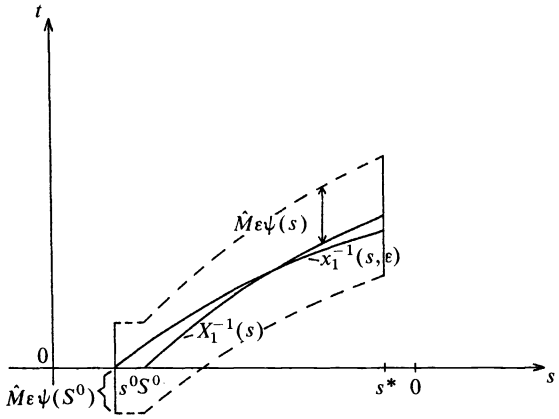


FIG. 1

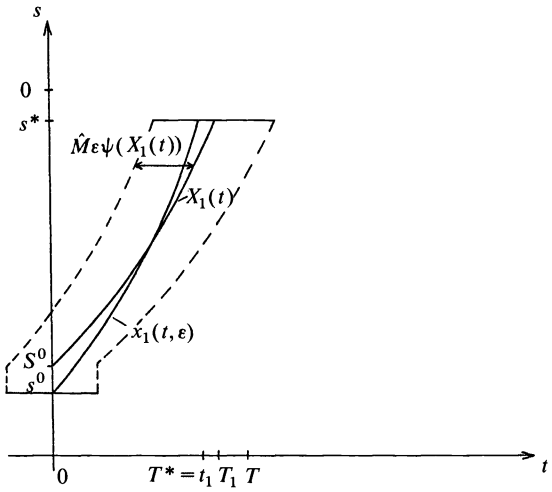


FIG. 2

The time T was defined by $T=X_1^{-1}(0)$. Moreover let

$$T_1(\varepsilon):=X_1^{-1}(s^*),$$
$$t_1(\varepsilon):=x_1^{-1}(s^*,\varepsilon),$$
$$T^*(\varepsilon):=\min(T_1,t_1).$$

It can easily be seen that there is $d>0$ such that

$$T^* \geqq T - d\varepsilon^{q*}.$$

Thus, for the functions $x_1(t,\varepsilon)$, $X_1(t)$ we obtain the situation sketched in Fig. 2. We now want to derive an estimate for $|x_1(t,\varepsilon)-X_1(t)|$ out of this picture. $X_1(t)$ provides a diffeomorphism from $[0,T]$ onto $[S^0,0]$ with inverse X_1^{-1}.

We first consider the lower boundary of the tube (see Fig. 3) and we put

$$r(t,\varepsilon) := \begin{cases} X_1(t) - X_1(\hat{t}(t,\varepsilon)), & T_0^* \leqq t \leqq T_1, \\ X_1(t) - s^0, & 0 \leqq t < T_0^*. \end{cases}$$

$r(t,\varepsilon)$ is a positive, continuous function on both intervals.

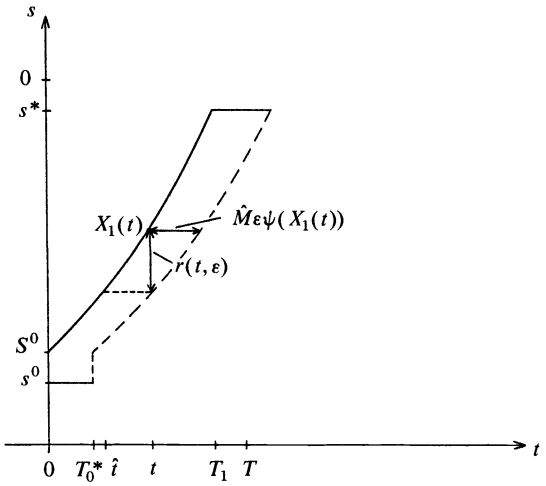


FIG. 3

The mean value theorem implies that

$$r(t,\varepsilon) = \dot{X}_1(\tau)\hat{M}\varepsilon\psi(X_1(\hat{t})) \quad \text{for some } \tau \in [\hat{t},t], t \in [T_0^*, T_1].$$

Hence, since $\psi(X_1(\hat{t})) < \psi(X_1(t))$ for $\hat{t} < t$, we have

$$r(t,\varepsilon) \leqq \tilde{M}\varepsilon\psi(X_1(t)) \quad \text{for } t \in [T_0^*, T_1].$$

For $t \in [0, T_0^*)$ we have

$$\begin{aligned} r(t,\varepsilon) &\leqq r(T_0^*,\varepsilon) + S^0 - s^0 \\ &\leqq K\varepsilon\psi(S^0) + O(\varepsilon) \leqq \tilde{K}\varepsilon\psi(S^0) \leqq \tilde{K}\varepsilon\psi(X_1(t)). \end{aligned}$$

Hence, there is $M_l > 0$ such that

$$r(t,\varepsilon) \leqq M_l\varepsilon\psi(X_1(t)) \quad \text{for } t \in [0, T_1].$$

The upper part of the tube (see Fig. 4) is slightly more difficult.

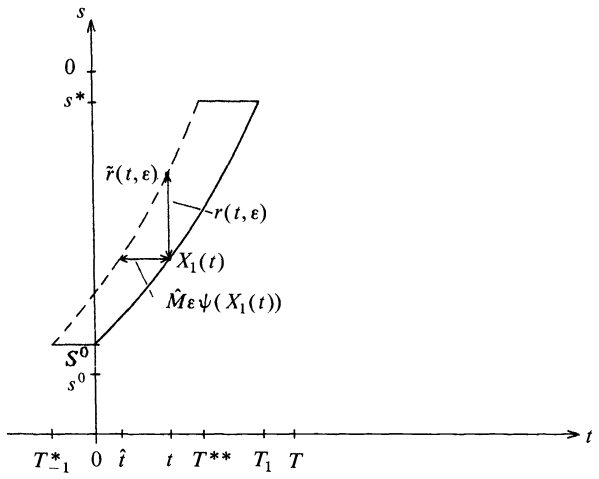


FIG. 4

The function $\tilde{r}(t,\varepsilon)$ (the upper boundary of the tube) is C^1 and increasing for $t\in[T^*_{-1}, T^{**}]$. This follows from the fact that the function $\tilde{r}^{-1}(s,\varepsilon)$ (the lower boundary of the tube in Fig. 1) is given by

$$\tilde{r}^{-1}(s,\varepsilon)=X_1^{-1}(s)-\hat{M}\varepsilon\psi(s) \quad \text{for } s\in[S^0,s^*],$$

and hence C^1 and increasing there with positive derivative. Therefore

$$r(t,\varepsilon):=\tilde{r}(t,\varepsilon)-\tilde{r}(\hat{t},\varepsilon)=\dot{\tilde{r}}(\tau,\varepsilon)\hat{M}\varepsilon\psi(X_1(t)) \quad \text{for some } \tau\in[\hat{t},t],\ t\in[0,T^{**}].$$

For $t\in[T^{**},T_1]$ we have for some $\tau<T^{**}$

$$\begin{aligned} r(t,\varepsilon)&\leqq r(T^{**},\varepsilon)=\dot{\tilde{r}}(\tau,\varepsilon)\hat{M}\varepsilon\psi(X_1(T^{**}))\\ &\leqq\dot{\tilde{r}}(\tau,\varepsilon)\hat{M}\varepsilon\psi(X_1(t)).\end{aligned}$$

Hence, since $\dot{\tilde{r}}(\tau,\varepsilon)$ is bounded for all possible τ there is $M_u>0$ such that

$$r(t,\varepsilon)\leqq M_u\varepsilon\psi(X_1(t)) \quad \text{for } t\in[0,T_1].$$

And we finally obtain that

$$|x_1(t,\varepsilon)-X_1(t)|<M_1\varepsilon\psi(X_1(t)) \quad \text{for } t\in[0,T^*]. \tag{7}$$

Since $X_1(T)=0$ and $\dot{X}_1(T)>0$ we have

$$X_1(t)=-M(T-t)+o(T-t), \qquad (T-t)\to 0^+.$$

Therefore, we finally obtain for T small enough (which can always be achieved by taking $|x_1^0|$ small enough) the following estimate:

$$|x_1(t,\varepsilon)-X_1(t)|<K_1\varepsilon\psi(T-t) \quad \text{for } t\in[0,T-d\varepsilon^{q*}].$$

For the remaining x-components we get from this result and again by means of [1, Thm. 2]:

$$\begin{aligned}|\bar{x}(t,\varepsilon)-\overline{X}(t)| &= |\bar{u}(x_1(t,\varepsilon),\varepsilon)-\overline{U}(X_1(t))| \\ &\leqq |\bar{u}(x_1(t,\varepsilon),\varepsilon)-\bar{u}(X_1(t),\varepsilon)|+|\bar{u}(X_1(t),\varepsilon)-\overline{U}(X_1(t))| \\ &< K|x_1(t,\varepsilon)-X_1(t)|+c\varepsilon\psi(X_1(t)) \\ &< \overline{K}\varepsilon\psi(T-t), \qquad t\in[0,T-d\varepsilon^{q*}].\end{aligned}$$

We now consider the y_1-component which is the most delicate one.

$$\begin{aligned}|y_1(t,\varepsilon)-Y_1(t)| &= |v_1(x_1(t,\varepsilon),\varepsilon)-V_1(X_1(t))| \\ &\leqq |v_1(x_1(t,\varepsilon),\varepsilon)-V_1(x_1(t,\varepsilon))|+|V_1(x_1(t,\varepsilon))-V_1(X_1(t))|.\end{aligned}$$

The first term can again be estimated by [1, Thm. 2]:

$$|v_1(x_1(t,\varepsilon),\varepsilon)-V_1(x_1(t,\varepsilon))|<c\varepsilon(-x_1(t,\varepsilon))^{-q_1}.$$

For the second term we obtain for each $t\in[0,T^*]$ by the mean-value theorem

$$|V_1(x_1)-V_1(X_1)|=|V_1'(x_1^*)||x_1-X_1|$$

for some x_1^* lying between x_1 and X_1. From [1] we know that $V_1'(s)=O((-s)^{\beta_1})$, $\beta_1=\min(\alpha-1,0)$, for $s\in[s^0,0)$. Hence

$$|V_1'(x_1^*)|\leqq N(-x_1^*)^{\beta_1}$$

and, since (7) implies that there is $\hat{N}>0$ such that for $\nu<0$

$$(-x_1(t,\varepsilon))^{\nu}\leqq\hat{N}(-X_1(t))^{\nu} \quad \text{for } t\in[0,T^*]$$

and ε sufficiently small, we get

$$\begin{aligned}|V_1(x_1(t,\varepsilon))-V_1(X_1(t))| &\leqq \tilde{N}\varepsilon(-X_1(t))^{\beta_1}\psi(X_1(t)) \\ &< N_1\varepsilon(-X_1(t))^{-q_1}, \qquad t\in[0,T^*].\end{aligned}$$

Therefore

$$\begin{aligned}|y_1(t,\varepsilon)-Y_1(t)| &< \hat{C}\varepsilon(-X_1(t))^{-q_1} \\ &< C_1\varepsilon(T-t)^{-q_1} \quad \text{for } t\in[0,T-d\varepsilon^{q*}].\end{aligned}$$

In a similar way, it can be shown that for ε small enough

$$|\bar{y}(t,\varepsilon)-\overline{Y}(t)|<\overline{C}\varepsilon\hat{\psi}(T-t) \quad \text{for } t\in[0,T-d\epsilon^{q*}],$$

where

$$\hat{\psi}(s):=\begin{cases}\log(-s)^{-1}, & q_1=1,\\ (-s)^{\bar{\alpha}-q_1}, & q_1>1.\end{cases}$$

This completes the proof of Theorem 1. □

3. Application to the nerve impulse equations of E. C. Zeeman. In [5] Zeeman suggests the following qualitative model for the local nerve impulse

$$\begin{aligned} \dot{x}_1 &= -1-x_2, \\ \dot{x}_2 &= -2x_2-2y, \\ \varepsilon\dot{y} &= -\left(x_1+x_2y+y^3\right), \end{aligned} \tag{8}$$

where x_1 corresponds to the potential of the membrane surrounding the axon of the neuron, and x_2 and y are correlated with the permeabilities of the membrane to potassium and sodium ions, respectively. Putting $\varepsilon=0$ in (8) we obtain the reduced system

$$\begin{aligned} \dot{x}_1 &= -1-x_2 =: f_1(x_1,x_2,y), \\ \dot{x}_2 &= -2x_2-2y =: f_2(x_1,x_2,y), \\ 0 &= -\left(x_1+x_2y+y^3\right) =: g(x_1,x_2,y), \end{aligned} \tag{9}$$

which defines a flow on the surface M: $g(x_1,x_2,y)\equiv 0$. M is sketched in Fig. 5. The fold curves $\tilde{C}^+$ and $\tilde{C}^-$ on M are given by

$$g_y = -\left(x_2+3y^2\right)=0.$$

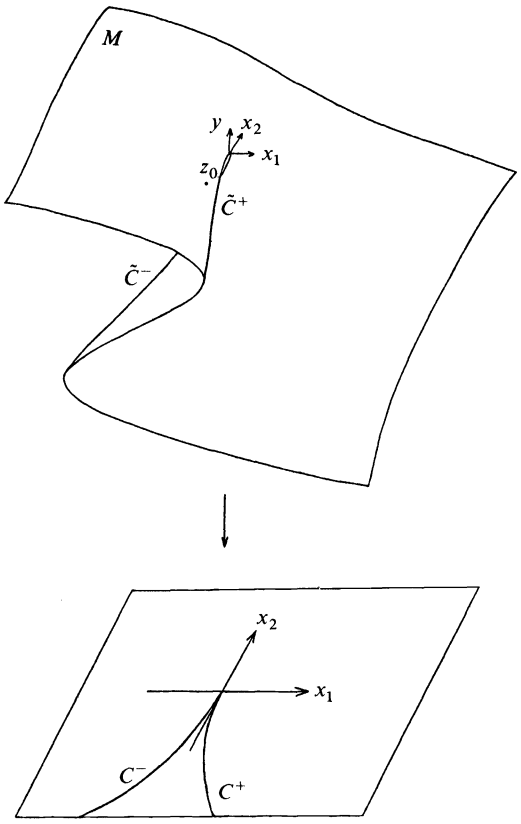


FIG. 5

When projected onto the (x_1, x_2)-plane they form a cusp defined by

$$27x_1^2 + 4x_2^3 = 0,$$

i.e. the curves C^+, C^- are given by

$$C^{\pm}: x_1 = \pm \frac{2}{3\sqrt{3}}(-x_2)^{3/2}, \qquad x_2 \leqq 0.$$

Outside the cusp M is single-sheeted, and inside it is 3-sheeted. Since the stability of these sheets as reduced manifolds of the system (8) is defined by $g_y < 0$ we find that the upper and lower sheet are stable (attractors) and the middle sheet is unstable (repellor).

The system (8) has one equilibrium point $z_0 = (0, -1, 1)$ which of course lies on M. To leading order in ε the eigenvalues of the Jacobian at z_0 are $-\frac{2}{\varepsilon}$, $e^{i2\pi/3}$, $e^{i4\pi/3}$, which in particular means that z_0 is a focus of the reduced flow on M. In Fig. 6 we have sketched the vector field of (8) on M. The curves $\tilde{A}, \tilde{B}, \tilde{C}$ on M are defined as follows:

$$\tilde{A}: f_2 = 0,$$
$$\tilde{B}: f_1 = 0,$$
$$\tilde{C}: g_y = 0 \text{ (fold curves)}.$$

It is easily verified that the projections A, B, C onto the (x_1, x_2)-plane look like indicated in Fig. 6.

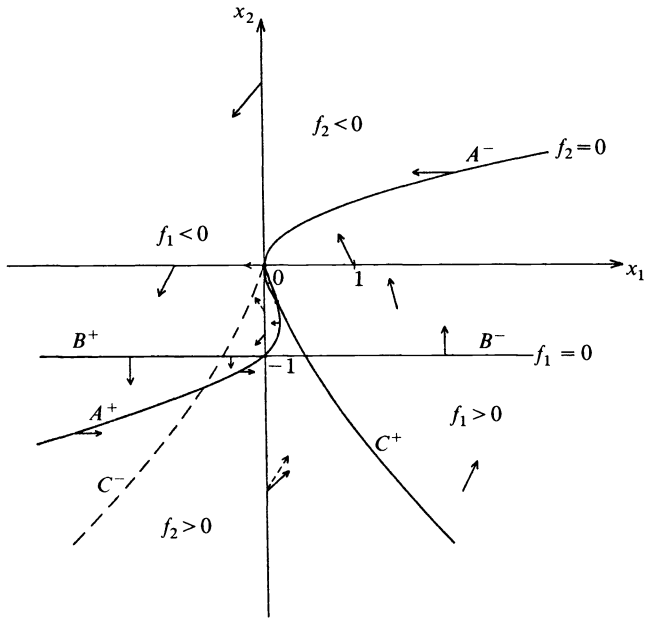


FIG. 6

In a finite distance away from M the vectors of the vector field of (8) are of length $O(1/\varepsilon)$ and hence for $\varepsilon \ll 1$ almost parallel to the y-axis. This means that the trajectory through such points drops into an ε-dependent vicinity of a stable sheet of M almost

instantaneously. It is plausible that the trajectory stays close to M as long as it does not encounter a point on a fold curve.

In the biological model, while no message is passing through the neuron, the membrane surrounding the axon is polarized, and the potential remains constant representing the stable equilibrium. As a message moves along the axon it triggers off a rapid depolarization of the membrane which causes the potential to increase suddenly. When the message has passed, the membrane repolarizes slowly, representing the smooth return to the equilibrium.

In the mathematical model (8), what happens if a trigger increases x_1 away from the equilibrium point z_0? A phase plane discussion (see Fig. 6) suggests that there are essentially two cases

(a) $x_1^0 \leqq \frac{2}{3\sqrt{3}}$,

(b) $x_1^0 > \frac{2}{3\sqrt{3}}$.

The value $x_1^0 = 2/3\sqrt{3}$ is a threshold for the trigger. In the case (a) the trajectory through $(x_1^0, -1, 1)$ will drop onto the upper sheet of M and in the case (b) onto the lower sheet of M. Then, the following situations may arise, depending on the size of x_1^0:

(A) (i) The trajectory spirals back into the equilibrium.
(ii) It encounters a point on the fold curve $\tilde{C}^+$ and drops down onto the lower sheet of M ($\rightarrow(B)$).

(B) (i) It slowly returns to the upper sheet of M and spirals into the equilibrium.
(ii) The trajectory encounters a point on the fold curve $\tilde{C}^-$ and drops onto the upper sheet of M and spirals into the equilibrium.

The case (B)(i) of course is the one that describes the biological model as sketched above. We, however, want to consider the case A(ii), B(i) which contains, from a mathematical point of view, all possible difficulties lying in this (mathematical) model. And we want to show that a trajectory of (8) is indeed approximated by jumps into an ε-dependent neighborhood of a stable sheet of M and trajectories (on M) of the reduced system (9). To be even more precise, we are going to prove rigorously that a trajectory of (8) (as well as the corresponding solution) is approximated by trajectories (solutions) of a sequence of reduced systems. Three different situations may arise:

—the reduced trajectory lies on a stable sheet of M,
—it encounters a point on a fold curve of M,
—it is "perpendicular" to M.

Hence without loss of generality we may consider a trajectory of (8) that starts in a neighborhood of a point on a fold curve. For convenience, we take the point $z^1 := (2, -3, 1)$ on $\tilde{C}^+$.

Let Γ^+ be the reduced trajectory of (9) through z^1 lying on the upper (stable) sheet of M. It is well defined for $x_1 \leqq 2$ in a neighborhood of z^1. Moreover, let $(X_1(t), X_2(t), Y(t))$ be the corresponding solution of (9) with initial conditions (at $t=0$) $Z^0 := (X_1^0, X_2^0, Y^0)$ on Γ^+ sufficiently close to z^1. It exists for $t \in (T_-, T)$, where $T_- < 0$ and $T > 0$ such that $(X_1(T), X_2(T), Y(T)) = z^1$.

We introduce local coordinates near z^1 by means of

$$x_1 = 2 + 2\hat{s} - 4x,$$
$$x_2 = -3 + 4\hat{s} + 2x,$$
$$y = 1 + \hat{y}.$$

Then the transformed equation (8) satisfies A1 and M is locally given by

$$\hat{s}=h(x,\hat{y})=-\frac{-2x+2x\hat{y}+3\hat{y}^2+\hat{y}^3}{6+4\hat{y}}.$$

Moreover, the fold curve $\tilde{C}^+$ has the (local) representation $(h(x,\varphi(x)),x,\varphi(x))$ for $x\in U(0)$, where the smooth function $\varphi(x)$ is such that

$$h_{\hat{y}}(x,\varphi(x))=0 \quad \text{for } x\in U(0),$$

with $\varphi(0)=0$.

In order that A3 holds we need one more transformation

$$\hat{s}=s+h(x,\varphi(x)), \qquad \hat{y}=v+\varphi(x).$$

In these new coordinates $\tilde{C}^+$ has the (local) representation $(0,x,0)$, $x\in U(0)$, and (8) now reads

$$\begin{aligned}
\dot{s}&=1-2s+\rho(x,v),\\
\dot{x}&=-\frac{1}{5}(v+\varphi(x)),\\
\varepsilon\dot{v}&=G(s,x,v)+\varepsilon\kappa(x,v)
\end{aligned} \tag{10}$$

where the smooth functions ρ, κ vanish at $(x,v)=(0,0)$ and

$$G(s,x,v)=-\left[(6+4\varphi(x))s+4sv+(3+3\varphi(x))v^2+v^3\right].$$

Therefore, M in these new coordinates is locally given by

$$-s=\frac{(3+3\varphi(x))v^2+v^3}{6+4\varphi(x)+4v}.$$

Taking the positive square root on both sides we obtain an equation in $\sqrt{-s},x,v$ that satisfies the IFT. Hence, we find that there is a continuous function $\phi(s,x)$ defined for $s\in(s_1,0]$, $x\in\Omega(0)$, which is in $C^1((s_1,0)\times\Omega)$ and has a continuous partial derivative with respect to x also for $s=0$, and which satisfies

$$\begin{aligned}
&G(s,x,\phi(s,x))=0 \quad \text{for } s\in(s_1,0],\ x\in\Omega,\\
&\phi(0,0)=0,
\end{aligned}$$

and

$$\begin{aligned}
&\phi(s,x)=\sqrt{-2s}+h.o.t.,\\
&\phi_s(s,x)=-\frac{1}{\sqrt{-2s}}+h.o.t., \qquad s\to 0^-,\quad |x|\to 0.
\end{aligned}$$

Hence, the upper (stable) sheet of M is locally given by $v=\phi(s,x)$ and A3 is satisfied with $\alpha=\frac{1}{2}$.

Let us now consider the following initial conditions to (10)

$$\begin{aligned}
s(0,\varepsilon)&=s^0(\varepsilon)=S^0+O(\varepsilon),\\
x(0,\varepsilon)&=x^0(\varepsilon)=X^0+O(\varepsilon),\\
v(0,\varepsilon)&=v^0(\varepsilon)=V^0+O(\varepsilon),
\end{aligned} \tag{11}$$

where (S^0, X^0, V^0) is the point Z^0 in the new coordinates, i.e. we start in an ε-dependent neighborhood of the reduced trajectory Γ^+ and close to the point z^1 on $\tilde{C}^+$ (such that the initial point lies in the s- and x-domain considered above). Then A4 is satisfied, and from

$$G_v(s,x,v) = -\left(4s + (6+6\varphi(x))v + 3v^2\right) \qquad \text{(which also implies A2)}$$

and by applying the estimates (27), (28) of [1] we find that A5 holds with $q = \frac{1}{2}$. In the same way we obtain $\beta = 0$, $\gamma = 1$.

Thus, all assumptions of Theorem 1 [1, Thm. 2] are satisfied, and we have the following result for the phase-plane trajectory of the IVP (10), (11):

$$\begin{aligned} &|x(s,\varepsilon) - X(s)| < c\varepsilon \log(-s)^{-1}, \\ &|v(s,\varepsilon) - V(s)| < c\varepsilon(-s)^{-1}, \end{aligned} \qquad s \in [s^0(\varepsilon), -c\varepsilon^{2/3}] \tag{12}$$

where $(X(s), V(s))$ is the trajectory of the reduced problem (the reduced trajectory Γ^+ in the new coordinates). Going back to the original variables we have for the corresponding solution of (8)

$$\begin{aligned} &|x_1(t,\varepsilon) - X_1(t)| < C\varepsilon \log(T-t)^{-1} &&< \hat{C}\varepsilon \log \varepsilon^{-1}, \\ &|x_2(t,\varepsilon) - X_2(t)| < C\varepsilon \log(T-t)^{-1} && \qquad\qquad t \in [0, T - C\varepsilon^{2/3}]. \\ &|y(t,\varepsilon) - Y(t)| < C\varepsilon(T-t)^{-1} &&\leqq \varepsilon^{1/3}, \end{aligned} \tag{13}$$

Hence, the reduced solution (trajectory) ceases to be an approximation in an ε-dependent neighborhood of the point z^1 where the stability breaks down. Guided by the algorithm presented in [2] we introduce the following shift scaling transformation

$$\begin{aligned} x_1 &= 2 + \varepsilon^{2/3} u_1, \\ x_2 &= -3 + \varepsilon^{2/3} u_2, \\ y &= 1 + \varepsilon^{1/3} q, \\ t &= T + \varepsilon^{2/3} \sigma, \end{aligned}$$

which means a blowing up of that neighborhood and takes (8) into

$$\begin{aligned} \frac{du_1}{d\sigma} &= 2 - \varepsilon^{2/3} u_2, \\ \frac{du_2}{d\sigma} &= 4 - 2\varepsilon^{1/3} q - 2\varepsilon^{2/3} u_2, \\ \frac{dq}{d\sigma} &= -\left(u_1 + u_2 + 3q^2\right) - \varepsilon^{1/3}\left(u_2 q + q^3\right). \end{aligned} \tag{14}$$

Let $(u_1(\sigma,\varepsilon), u_2(\sigma,\varepsilon), q(\sigma,\varepsilon))$ be the solution of (14) that satisfies

$$\begin{aligned} u_1(-\hat{\sigma},\varepsilon) &= \varepsilon^{-2/3}\left[x_1(T - \hat{\sigma}\varepsilon^{2/3},\varepsilon) - 2\right], \\ u_2(-\hat{\sigma},\varepsilon) &= \varepsilon^{-2/3}\left[x_2(T - \hat{\sigma}\varepsilon^{2/3},\varepsilon) + 3\right], \\ q(-\hat{\sigma},\varepsilon) &= \varepsilon^{-1/3}\left[y(T - \hat{\sigma}\varepsilon^{2/3},\varepsilon) - 1\right], \end{aligned} \tag{15}$$

for some $\hat{\sigma} \geqq C$. A solution of the reduced system corresponding to (14) satisfies

$$\begin{aligned} &U_1(\sigma) = 2\sigma + U_1^0, \qquad U_2(\sigma) = 4\sigma + U_2^0, \\ &\dot{Q} = -6\sigma - 3Q^2 - \left(U_1^0 + U_2^0\right). \end{aligned} \tag{16}$$

Taking $U_1^0 = U_2^0 = 0$ it can be shown that

$$\begin{aligned} |u_1(-\hat{\sigma},\varepsilon) - U_1(-\hat{\sigma})| &= O(\varepsilon^{1/3}\log\varepsilon^{-1}), \\ |u_2(-\hat{\sigma},\varepsilon) - U_2(-\hat{\sigma})| &= O(\varepsilon^{1/3}\log\varepsilon^{-1}), \end{aligned} \qquad \varepsilon \to 0.$$

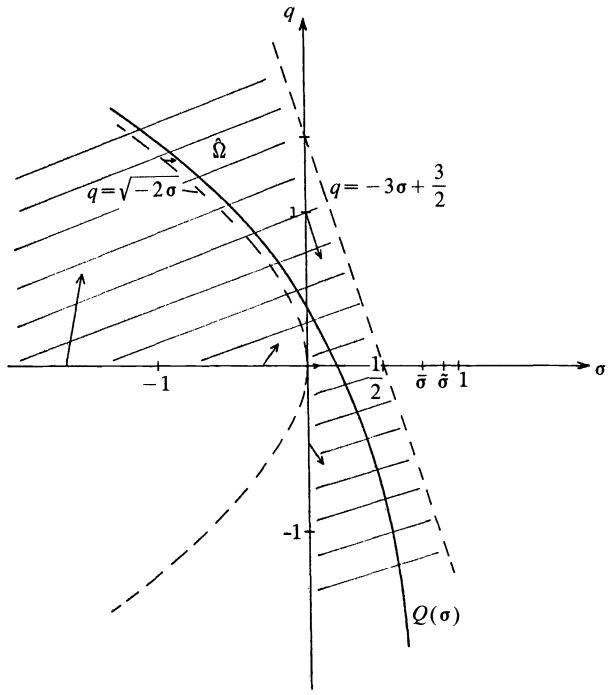


FIG. 7

Equation (16) is discussed in Fig. 7. The domain $\hat{\Omega}$ sketched there is invariant, i.e. every solution $Q(\sigma,\sigma_0,Q^0)$ with $(\sigma_0,Q^0)\in\hat{\Omega}$ stays in $\hat{\Omega}$ for all $\sigma\in(\sigma_0,\sigma_+)$, where (σ_-,σ_+), $\sigma_+ > 0$, is its maximal domain of existence. Moreover, it follows that $Q(\sigma,\sigma_0,Q^0)\to-\infty$ for $\sigma\to\sigma_+$. Since (13) implies

$$|q(-\hat{\sigma},\varepsilon) - \sqrt{2\hat{\sigma}}| < 1$$

we know that $q(-\hat{\sigma},\varepsilon)\in\hat{\Omega}$ for $\hat{\sigma}$ large enough. Hence, applying standard arguments (Gronwall lemma) we obtain that there is $0<\bar{\sigma}<\frac{5}{6}$ such that for ε sufficiently small the solution of (14) with initial conditions (15) exists at least for $\sigma\in[-\hat{\sigma},\bar{\sigma}]$ and

$$q(\bar{\sigma},\varepsilon) < -1. \tag{17}$$

Equation (16) can be transformed into Airy's equation which is linear of second order (see e.g. [2, Ex. 1]). Let $Q(\sigma)$ be the solution of (16) based on the first one of the two linearly independent solutions of Airy's equation. Then, it can be shown that

$$Q(\sigma) = \sqrt{-2\sigma} + O\big((-\sigma)^{-1}\big), \qquad \sigma\to-\infty$$

and

$$|q(-\hat{\sigma},\varepsilon)-Q(-\hat{\sigma})|<1 \quad \text{for } \hat{\sigma} \text{ large enough.}$$

Moreover, $Q(\sigma)$ has a pole at $\sigma=\tilde{\sigma}\cong 0.892$, and hence behaves as sketched in Fig. 7. Thus, by means of the same arguments as before we obtain for our solution $(x_1(t,\varepsilon),x_2(t,\varepsilon),y(t,\varepsilon))$ of (8) that for every $c_2\in(0,\tilde{\sigma}-\bar{\sigma}]$ there exists $M>0$ such that

$$\begin{aligned} &|x_1(t,\varepsilon)-\tilde{X}_1(t)|<M\varepsilon\log\varepsilon^{-1},\\ &|x_2(t,\varepsilon)-\tilde{X}_2(t)|<M\varepsilon\log\varepsilon^{-1}, \qquad t\in\left[T-\hat{\sigma}\varepsilon^{2/3},T+(\tilde{\sigma}-c_2)\varepsilon^{2/3}\right],\\ &|y(t,\varepsilon)-\tilde{Y}(t)|<M\varepsilon^{1/3}, \end{aligned} \tag{18}$$

where $(\tilde{X}_1(t),\tilde{X}_2(t),\tilde{Y}(t))$ is the reduced solution $(U_1(\sigma),U_2(\sigma),Q(\sigma))$ expressed in the original variables.

By means of this second approximation, we have passed the delicate point $z^1\in\tilde{C}^+$. We are still in an ε-dependent neighborhood, however.

Again we proceed as motivated in [2] and introduce the new shift scaling transformation

$$\begin{aligned} u_1&=2\tilde{\sigma}+\varepsilon^{1/3}v_1,\\ u_2&=4\tilde{\sigma}+\varepsilon^{1/3}v_2,\\ q&=\varepsilon^{-1/3}w,\\ \sigma&=\tilde{\sigma}+\varepsilon^{1/3}\tau, \end{aligned}$$

that takes (14) into

$$\begin{aligned} \frac{dv_1}{d\tau}&=2-4\varepsilon^{2/3}\tilde{\sigma}-\varepsilon v_2,\\ \frac{dv_2}{d\tau}&=4-2w-8\varepsilon^{2/3}\tilde{\sigma}-2\varepsilon v_2,\\ \frac{dw}{d\tau}&=-3w^2-w^3-\varepsilon^{2/3}(6\tilde{\sigma}+4\tilde{\sigma}w)-\varepsilon(v_1+v_2+v_2w). \end{aligned} \tag{19}$$

Let $(v_1(\tau,\varepsilon),v_2(\tau,\varepsilon),w(\tau,\varepsilon))$ be the solution of (19) satisfying

$$\begin{aligned} v_1(-\varepsilon^{-1/3}c_2,\varepsilon)&=\varepsilon^{-1}\left[x_1\left(T+\varepsilon^{2/3}(\tilde{\sigma}-c_2),\varepsilon\right)-(2+\varepsilon^{2/3}2\tilde{\sigma})\right],\\ v_2(-\varepsilon^{-1/3}c_2,\varepsilon)&=\varepsilon^{-1}\left[x_2\left(T+\varepsilon^{2/3}(\tilde{\sigma}-c_2),\varepsilon\right)+3-\varepsilon^{2/3}4\tilde{\sigma}\right],\\ w(-\varepsilon^{-1/3}c_2,\varepsilon)&=y\left(T+\varepsilon^{2/3}(\tilde{\sigma}-c_2),\varepsilon\right)-1. \end{aligned}$$

From (17) we know that $w(-\varepsilon^{-1/3}c_2,\varepsilon)<-\varepsilon^{1/3}$. Typical solutions of the reduced w-equation are sketched in Fig. 8. A solution $W(\tau,W^0)$, $W(0,W^0)=W^0$, with $W^0\in(-3,0)$ exists for $\tau\in(-\infty,\infty)$ and approaches 0 geometrically for $\tau\to-\infty$ and -3 exponentially for $\tau\to+\infty$. By using "phase-plane" arguments in the (w,τ)-plane and Gronwall type results it can be shown that $(v_1(\tau,\varepsilon),v_2(\tau,\varepsilon),w(\tau,\varepsilon))$ exists for $\tau\in[-c_2\varepsilon^{-1/3},\varepsilon^{-1/3}]$. Moreover, there is $c_3\in(0,1)$ such that

$$\begin{aligned} v_1(\tau,\varepsilon)&=O(\tau),\\ v_2(\tau,\varepsilon)&=O(\tau), \qquad \tau\in\left[c_3\varepsilon^{-1/3},\varepsilon^{-1/3}\right].\\ w(\tau,\varepsilon)&=-3+O(\varepsilon^{2/3}), \end{aligned}$$

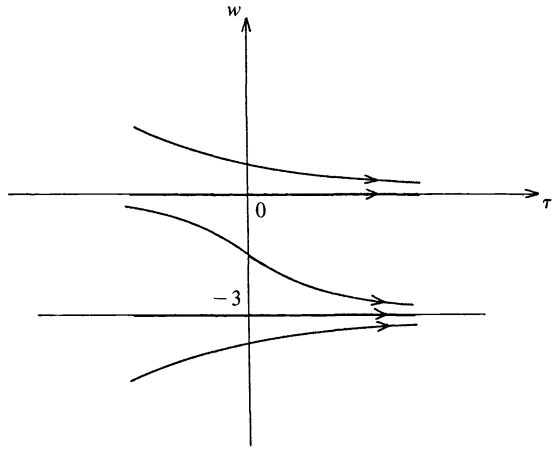


FIG. 8

Thus in the original variables we have

$$
\begin{aligned}
&x_1\big(T+\varepsilon^{2/3}(\tilde{\sigma}+1),\varepsilon\big)=2+O(\varepsilon^{2/3}),\\
&x_2\big(T+\varepsilon^{2/3}(\tilde{\sigma}+1),\varepsilon\big)=-3+O(\varepsilon^{2/3}),\\
&y\big(T+\varepsilon^{2/3}(\tilde{\sigma}+1),\varepsilon\big)=-2+O(\varepsilon^{2/3}).
\end{aligned}
\tag{20}
$$

The point $z^2:=(2,-3,-2)$ lies on the lower (stable) sheet of the surface M. Hence, we have proved up to now that (and how) the trajectory of our solution $(x_1(t,\varepsilon),x_2(t,\varepsilon),y(t,\varepsilon))$ of (8) has dropped off the upper sheet of M down to the lower one. It drops down almost instantaneously (time $O(\varepsilon^{2/3})$) and almost vertically ($O(\varepsilon^{2/3})$-neighborhood of $(x_1,x_2)=(2,-3)$).

We again consider the original system (8) now together with the initial conditions (20). By simple phase plane arguments (compare Fig. 6) it can be shown that the trajectory of the corresponding reduced problem (lying on the lower sheet of M) crosses the curve $\tilde{A}^-$, follows $\tilde{A}^-$ into a neighborhood of the origin where it leaves $\tilde{A}^-$ now being on the upper sheet of M. Then it crosses $\tilde{A}^+$ and follows $\tilde{A}^+$ into a neighborhood of the point $(0,-1,1)$ and finally spirals into this point on M which is a stationary point (focus) of the reduced system. Hence, the corresponding reduced solution $(\bar{X}_1(t),\bar{X}_2(t),\bar{Y}(t))$ exists for all $t\geqq T+\varepsilon^{2/3}(\tilde{\sigma}+1)$ and since the trajectory stays on the stable sheets of M we may apply the Tikhonov theorem (cf. [1]) on every finite t-interval $[T+\varepsilon^{2/3}(\tilde{\sigma}+1),c]$. Taking also into account that $z_0=(0,-1,1)$ is an asymptotically stable equilibrium solution of the full system (8) we thus obtain for ε small enough the following result on the unbounded t-interval:

$$
\begin{aligned}
&|x_1(t,\varepsilon)-\bar{X}_1(t)|<K\varepsilon^{2/3}\\
&|x_2(t,\varepsilon)-\bar{X}_2(t)|<K\varepsilon^{2/3}\quad \text{for every } t\geqq T+\varepsilon^{2/3}(\tilde{\sigma}+1).\\
&|y(t,\varepsilon)-\bar{Y}(t)|<K\varepsilon^{2/3}
\end{aligned}
$$

This was the final step. Summarizing, we have proved that the solution $(x_1(t,\varepsilon),x_2(t,\varepsilon),y(t,\varepsilon))$ we have considered of the system (8) exists for all $t\geqq 0$ and its trajectory behaves as sketched in Fig. 9. The motion is slow as long as the trajectory is

close to a stable sheet of M. The jump from the upper sheet of M to the lower one happens in a fast time scale.

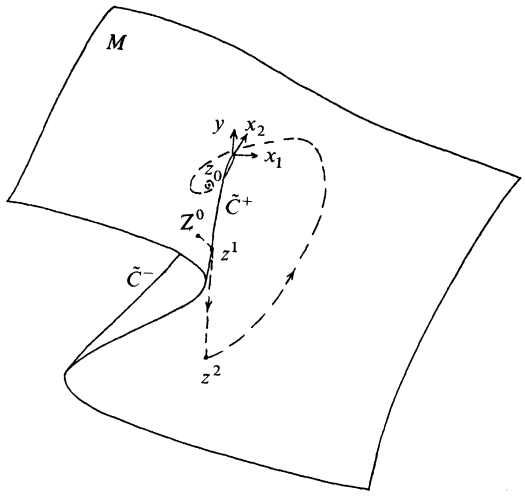


FIG. 9

4. The case "$x_1 > 0$". This case is characterized by the fact that the flow along a stable reduced trajectory of (2) leads away from the origin, the point where the stability assumption is violated (see [1]). Also in this case, the estimates for the trajectories of (1) given in [1, Thm. 3] carry over to the solutions and in a way completely analogous to the proof of Theorem 1. For completeness, we state these estimates, but without formulating the precise Theorem 2 corresponding to [1, Thm. 3] and without proof.

Let $(\bar{u}(x_1,\varepsilon), v(x_1,\varepsilon))$ and $(\bar{U}(x_1), V(x_1))$ be trajectories of (1) and (2), respectively, that satisfy [1, Thm. 3] and let c, δ and q^* be the quantities defined there. Moreover, let $(x(t,\varepsilon), y(t,\varepsilon))$ and $(X(t), Y(t))$ be the corresponding solutions obtained by introducing the time t in the way (compare §§1, 2)

$$x_1(0,\varepsilon) = s^*(\varepsilon), \qquad X_1(0) = S^*(\varepsilon)$$

with

$$s^*(\varepsilon) = c\varepsilon^{q^*} \quad \text{and} \quad |s^* - S^*| = O(\varepsilon^{\delta}).$$

Then, if the parameter q_1 in [1, Thm. 3] is taken such that $q_1 > 1 - \hat{\alpha}$ (i.e. $q^* < \delta$), there are positive constants C and c_1 such that the following estimates hold for T_1 and ε small enough:

$$\begin{aligned}
&|x(t,\varepsilon) - X(t)| < C\varepsilon^{\delta}, \\
&|y_1(t,\varepsilon) - Y_1(t)| < C\varepsilon^{\delta}\left(c_1\varepsilon^{q^*} + t\right)^{-q_1}, \\
&|\bar{y}(t,\varepsilon) - \bar{Y}(t)| < C\varepsilon^{\delta}\left(c_1\varepsilon^{q^*} + t\right)^{\bar{\omega}-q_1},
\end{aligned}$$

for $t \in [0, T_1]$.

Examples for this case "$x_1 > 0$" (as well as for "$x_1 < 0$") are the situations treated in [3], [4].

REFERENCES

[1] K. NIPP, *Breakdown of stability in singularly perturbed autonomous systems—I. Orbit equations*, SIAM J. Math. Anal., 17 (1986), pp. 512–532.

[2] _____, *An algorithmic approach to singular perturbation problems in* ODE's *with an application to the Belousov–Zhabotinskii reaction*, Ph.D. thesis, ETH Zurich, No. 6643, 1980.

[3] N. R. LEBOVITZ AND R. J. SCHAAR, *Exchange of stabilities in autonomous systems*, Stud. Appl. Math., 54 (1975), pp. 229–260.

[4] _____, *Exchange of stabilities in autonomous systems—II. Vertical bifurcation*, Stud. Appl. Math., 56 (1977), pp. 1–50.

[5] E. C. ZEEMAN, *Differential equations for the heartbeat and nerve impulse*, in Dynamical Systems, M. M. Peixoto, ed., Symposium Bahia, Brazil, 1971.

SIAM J. MATH. ANAL.
Vol. 17, No. 5, September 1986

004

ON THE GLOBAL REPRESENTATION OF THE SOLUTIONS OF SECOND-ORDER LINEAR DIFFERENTIAL EQUATIONS HAVING AN IRREGULAR SINGULARITY OF RANK ONE IN ∞ BY SERIES IN TERMS OF CONFLUENT HYPERGEOMETRIC FUNCTIONS*

THEO KURTH† AND DIETER SCHMIDT†

Abstract. A general second-order linear differential equation having an irregular singularity of rank one in ∞ is considered. It is shown that the solutions of this equation can be represented by series in terms of confluent hypergeometric functions which describe the full analytic behavior at the singular point ∞.

Key words. global representations, singular ODEs, series, confluent hypergeometric functions

AMS(MOS) subject classifications. Primary 34A20, 34E05, 33A30

Introduction. In the theory of complex ordinary differential equations a study of the global behavior of the solutions is one of the most interesting and difficult problems. A specific problem of this kind consists in finding global representations for the solutions which describe their full analytic behavior. In the present paper we want to treat this question for the case of the general second-order linear differential equation (DE) having an irregular singular point of rank one in ∞, which is of the form

$$y''+f(z)y'+g(z)y=0,$$

where f and g are holomorphic in ∞. There is a very extensive literature on this DE since the fundamental papers of Horn [5] in 1908 and Birkhoff [1] in 1913. (See e.g. [12], [6], [7], [8], [4].)

By a transformation of variables the above DE takes the following *normal form*

$$y''+\left(\frac{1}{z}+\frac{a(z)}{z^2}\right)y'+\left(1+\frac{2i\kappa}{z}+\frac{b(z)}{z^2}\right)y=0, \tag{0.1}$$

where κ is a complex constant and $a(z)$, $b(z)$ are functions being holomorphic in a full neighborhood $r<|z|\leqq\infty$ of infinity. We shall restrict our considerations to the DE (0.1).

It is well known that the solutions of the DE (0.1) are holomorphic functions on the Riemann surface of $\ln z$ over $r<|z|<\infty$, which possess in each sector $|\arg z-\pi/2-n\pi|<\pi/2$, $(n\in\mathbb{Z})$ a characteristic asymptotic behavior as $|z|\to\infty$. (See e.g. [11], pp. 232–236.) Our main aim is to obtain for these solutions global representations which are valid on the whole Riemann surface and which completely reflect the asymptotic behavior for $|z|\to\infty$ as well as the transition behavior in the case of analytic continuation around ∞.

A central role in our consideration plays the special case of the DE (0.1) with $a(z)$ equal to 0 and $b(z)$ equal to a constant $-\mu^2$:

$$y''+\frac{1}{z}y'+\left(1+\frac{2i\kappa}{z}-\frac{\mu^2}{z^2}\right)y=0. \tag{0.2}$$

*Received by the editors August 21, 1984, and in revised form July 31, 1985
†Universität Gesamthochschule Essen, FB 6 Mathematik, D-4300 Essen 1, West Germany.

This DE is—up to a simple transformation of variables—the confluent hypergeometric DE. The solutions of this DE—the confluent hypergeometric functions—are very well known. Using a lot of these well known facts and taking advantage of the very close relationship between the DEs (0.1) and (0.2), we can show that the solutions of the DE (0.1) are to be represented by series in terms of confluent hypergeometric functions, and that these representations are global ones possessing all the desired properties. A very specific property of these series, which underlines their usefulness, is that essentially only one set of coefficients is needed for the expansion of different solutions of the DE (0.1).

An immediate cause for the present consideration results from the study of special functions of mathematical physics, where expansions of "higher" special functions in terms of "simple" special functions have played a central role at all times. It is just this field from which we have obtained essential stimulations. In this connection we should mention, above all, the corresponding considerations of Meixner and Schäfke in Chapter 3 of their book on Mathieu and spheroidal functions [9, pp. 289–300], where spheroidal functions are expanded in terms of Bessel and Hankel functions. As far as the whole organization and the analytic methods of this paper are concerned there is a close relationship with the paper [16] of the second author.

This paper is divided into two main sections. Section 1 contains preliminary results: in §1.1 and §1.2 we introduce the fundamental notations for the DEs (0.1) and (0.2) and give some elementary, respectively well-known statements. In §1.3 we make available a suitable version of a theorem of F. W. Schäfke on expansions in terms of Whittaker functions, which serves as a main tool in our analysis, and in the following we draw some important conclusions on the coefficients of such expansions. In §1.4 we prove a theorem on the asymptotic behavior of series in terms of confluent hypergeometric functions, which surely is interesting by itself. This theorem is a generalization of a corresponding result of Meixner and Schäfke [9, pp. 95–97]. Section 2 contains the main results on the representation of the solutions of the DE (0.1) in terms of confluent hypergeometric functions. In §2.1 we transform the DE (0.1) into sequence spaces. In §2.2 we prove the fundamental result concerning the asymptotic behavior of the expansion coefficients which ensures the global convergence of all series. Section 2.3 finally contains the definition and a detailed discussion of the solutions of the DE (0.1) by means of their series representations.

Three final remarks are to be made:

1) Our analysis of the DE (0.1) also furnishes the *existence* of asymptotic solutions of the DE (0.1), certainly under the use of the very strong expansion theorem in §1.3. Another method to obtain the main results of this paper is to base all considerations on the existence of asymptotic solutions and to use the principle of analytic equivalence. (See [6]–[8]).

2) All series representations in this paper are only valid in the case that the characteristic exponents ν of the DE (0.1) fulfill $2\nu \notin \mathbb{Z}$ ("*normal case*"). A treatment of the exceptional case $2\nu \in \mathbb{Z}$ in the frame of our methods would require a more recent expansion theorem of F. W. Schäfke in [15, pp. 134–135], which is based on the general results in 1.1.11 of [10].

3) In principle, it should be possible to extend the methods of this paper also to general DEs of higher rank, but considerable efforts are to be made: First, one has to find an adequate set of "special" DEs being analytic equivalent to the "general" DE. Secondly, one has to establish for these special DEs all the results corresponding to those for equation (0.2) which are contained in §§1.2, 1.3 and 1.4.

1. Preliminary results.

1.1. General remarks on the DE (0.1); notation. In the following let Ω denote the annulus $\{z\in\mathbb{C}\,|\,r<|z|<\infty\}$ and $\hat{\Omega}$ the Riemann surface of $\ln z$ over Ω.[1] The vector space of holomorphic functions in $\hat{\Omega}$ is then denoted by $\mathscr{H}(\hat{\Omega})$.

For $y\in\mathscr{H}(\hat{\Omega})$ we define

$$L_\kappa y := \left(z\frac{d}{dz}\right)^2 y+(z^2+2i\kappa z)y, \tag{1.1}$$

where $\kappa\in\mathbb{C}$, and

$$Gy := \frac{d}{dz}(g_1\cdot y)+g_2\cdot y, \tag{1.2}$$

where the coefficients $g_1 := a$ and $g_2 := b-da/dz$ are holomorphic functions in $\Omega\cup\{\infty\}$. L_κ and G are linear operators in $\mathscr{H}(\hat{\Omega})$, in terms of which the DE (0.1) then reads

$$Dy := L_\kappa y+Gy=0. \tag{D}$$

The solution space of (D), which we denote in the following by $\mathfrak{N}(D)$, is a two-dimensional subspace of $\mathscr{H}(\hat{\Omega})$.

The following remark is to be easily verified:

(1.3) *Let $y_1,y_2\in\mathfrak{N}(D)$. Then the Wronskian*

$$w[y_1,y_2] := y_1\cdot\frac{d}{dz}y_2-y_2\cdot\frac{d}{dz}y_1$$

has the form

$$w[y_1,y_2](z)=[y_1,y_2]\cdot z^{-1}\cdot\exp\left(\int_z^\infty g_1(\zeta)\zeta^{-2}d\zeta\right),$$

where $[y_1,y_2]\in\mathbb{C}$. $[\cdot,\cdot]$ is a nontrivial alternating bilinear form on $\mathfrak{N}(D)$.

Defining for $y\in\mathscr{H}(\hat{\Omega})$

$$(\phi y)(z) := y(e^{2\pi i}\cdot z)\qquad(z\in\hat{\Omega}), \tag{1.4}^2$$

we obtain a linear operator ϕ in $\mathscr{H}(\hat{\Omega})$ which maps $\mathfrak{N}(D)$ onto itself: $\phi(\mathfrak{N}(D))=\mathfrak{N}(D)$. The restriction $\phi|_{\mathfrak{N}(D)}$ is then called the "*monodromy operator of the DE (D)*". It is one of the fundamental quantities of the DE (D). Further, one defines $\nu\in\mathbb{C}$ to be a "*characteristic exponent of the DE (D)*", iff $\exp(2\pi i\nu)$ is an eigenvalue of the monodromy operator $\phi|_{\mathfrak{N}(D)}$. This means especially that there exists a nontrivial solution y of (D) belonging to

$$\mathfrak{U}_\nu := (\phi-e^{2\pi i\nu})^{-1}(\{0\})=\{z^\nu\cdot h\,|\,h:\Omega\to\mathbb{C}\text{ holomorphic}\}. \tag{1.5}$$

A solution of this kind is called "*Floquet solution of the DE (D) with respect to the characteristic exponent ν*".

Let the set of all characteristic exponents of (D) be denoted by Ξ. Ξ is then characterized by

$$\nu\in\Xi\Leftrightarrow\cos 2\pi\nu=\frac{1}{2}\operatorname{trace}\phi\Big|_{\mathfrak{N}(D)}. \tag{1.6}$$

---

[1] For the sake of briefness we use the same symbol z for the point $(r,\varphi)\in\hat{\Omega}$ as well as for its projection $re^{i\varphi}\in\Omega$.

[2] For $z=(r,\varphi)$, $x=(\rho,\psi)\in\hat{\Omega}$ let $z\cdot x := (r\cdot\rho,\varphi+\psi)\in\hat{\Omega}$.

Thus, for any $\nu \in \Xi$ we have $\Xi = (\nu + \mathbb{Z}) \cup (-\nu + \mathbb{Z})$. For the sake of completeness we give a short proof of (1.6), using the nontrivial alternating form $[\cdot, \cdot]$ in (1.3): Let y_1, y_2 be a fundamental set of solutions of (D) with $[y_1, y_2] = 1$. We then have

$$(*) \qquad \det \phi|_{\mathfrak{N}(D)} = [\phi y_1, \phi y_2], \qquad \operatorname{trace} \phi|_{\mathfrak{N}(D)} = [\phi y_1, y_2] + [y_1, \phi y_2].$$

Since the Wronskian $w[y_1, y_2]$ is holomorphic in Ω, we obtain

$$w[\phi y_1, \phi y_2] = \phi(w[y_1, y_2]) = w[y_1, y_2]$$

and therefore

$$(**) \qquad \det \phi|_{\mathfrak{N}(D)} = [\phi y_1, \phi y_2] = [y_1, y_2] = 1.$$

Since the eigenvalues $\lambda = \exp(2\pi i \nu)$ of $\phi \mid_{\mathfrak{N}(D)}$ are determined by

$$\lambda^2 - \lambda \cdot \operatorname{trace} \phi|_{\mathfrak{N}(D)} + \det \phi|_{\mathfrak{N}(D)} = 0,$$

we immediately obtain (1.6).

It should be observed, however, that the representation of $\operatorname{trace} \phi \mid_{\mathfrak{N}(D)}$ in (*) does not provide an elementary means of calculating the characteristic exponents of (D).

Finally it turns out to be convenient introducing for $\nu \in \mathbb{C}$ the operator

$$(1.7) \qquad Q_\nu := \frac{1}{2\pi i}(\phi - e^{2\pi i \nu}).$$

Obviously, for $\nu \in \mathbb{C}$ and $y \in \mathscr{H}(\hat{\Omega})$

$$(1.8.1) \qquad \frac{\sin 2\pi\nu}{\pi} y = Q_{-\nu} y - Q_\nu y.$$

Moreover, for $\nu \in \Xi$ and $y \in \mathfrak{N}(D)$

$$(1.8.2) \qquad Q_{\mp\nu} y \in \mathfrak{U}_{\pm\nu} \cap \mathfrak{N}(D).$$

In the "*normal case*" of the DE (D) where

$$2\nu \notin \mathbb{Z} \text{ for (one and hence for all) } \nu \in \Xi$$

the formula (1.8.1) thus yields a unique representation of each $y \in \mathfrak{N}(D)$ in terms of Floquet solutions.

1.2. Floquet and asymptotic solutions of the DE (0.2). In terms of the operator L_κ the DE (0.2) can be written in the form

$$(D_\mu) \qquad D_\mu y := L_\kappa y - \mu^2 y = 0,$$

where $\kappa, \mu \in \mathbb{C}$. By the transformations

$$y(z) = e^{\mp iz} \cdot z^\mu \cdot u(x), \qquad x = \pm 2iz,$$

this DE is equivalent to the confluent hypergeometric DE

$$xu'' + (c - x)u' - au = 0,$$

where $a = \frac{1}{2} \mp \kappa + \mu$, $c = 1 + 2\mu$.

Introducing a slightly modified Φ function

$$\tilde{\Phi}(a;\, c;\, x) := \sum_{n=0}^{\infty} \frac{(a)_n}{\Gamma(c+n)} \frac{x^n}{n!} \qquad (x \in \mathbb{C}),$$

where $(a)_n = \Gamma(a+n)/\Gamma(a) = a(a+1)\cdots(a+n-1)$, we obtain by

$$I_{\kappa,\mu}(z) := e^{-iz} \cdot z^{\mu} \cdot \tilde{\Phi}\left(\frac{1}{2} - \kappa + \mu;\, 1+2\mu;\, 2iz\right) \tag{1.9}$$

a solution of the DE (D_μ). $\tilde{\Phi}$ being holomorphic with respect to (x,a,c); the same is true for $I_{\kappa,\mu}$ with respect to (z,κ,μ). Obviously, $I_{\kappa,\mu}$ belongs to $\mathfrak{U}_\mu$ and is hence a Floquet solution of (D_μ), if $\neq 0$; this is especially true for $2\mu \notin \mathbb{Z}$.

Since the DE (D_μ) is invariant under the substitution $\mu \to -\mu$, we obtain by $I_{\kappa,-\mu}$ a second solution of (D_μ), which now belongs to $\mathfrak{U}_{-\mu}$ and is $\neq 0$ for $2\mu \notin \mathbb{Z}$. Therefore, in the case of $2\mu \notin \mathbb{Z}$, $I_{\kappa,\mu}$ and $I_{\kappa,-\mu}$ constitute a fundamental set of solutions.

Introducing the Ψ function (see e.g. [3, p. 255]), we obtain by

$$H^{\pm}_{\kappa,\mu}(z) := (2e^{\mp \pi i/2})^{1/2 \pm \kappa + \mu} \cdot e^{\pm iz} \cdot z^{\mu} \cdot \Psi\left(\frac{1}{2} \pm \kappa + \mu;\, 1+2\mu;\, 2e^{\mp \pi i/2} \cdot z\right) \tag{1.10}$$

solutions of the DE (D_μ), which possess the asymptotic expansions

$$H^{\pm}_{\kappa,\mu}(z) \sim e^{\pm iz} \cdot z^{-1/2 \mp \kappa} \sum_{m=0}^{\infty} \frac{(1/2 \pm \kappa + \mu)_m (1/2 \pm \kappa - \mu)_m}{m!} \cdot (\pm 2iz)^{-m} \tag{1.11}$$

as $|z| \to \infty$ within the sectors $|\arg z \mp \pi/2| < 3\pi/2$. Here—and also in the following—let the upper and lower signs always be taken simultaneously.

The different asymptotic behavior of $H^{+}_{\kappa,\mu}$ and $H^{-}_{\kappa,\mu}$ implies that these functions constitute a fundamental set of solutions of the DE (D_μ) without any restrictions to the values of κ and μ.

The functions $H^{\pm}_{\kappa,\mu}$ possess the following representations as Laplace integrals

$$H^{\pm}_{\kappa,\mu}(z) = e^{\pm iz} \cdot z^{1/2 \mp \kappa} \cdot \int_0^{\infty \cdot e^{-i\xi}} e^{-zt}\, {}_2F_1\left(\frac{1}{2} \pm \kappa + \mu,\, \frac{1}{2} \pm \kappa - \mu;\, 1;\, \pm \frac{t}{2i}\right) dt, \tag{1.12}$$

which are valid for $z \in \hat{\Omega}$ with $|\arg z - \xi| < \pi/2$ where $\xi \in \mathbb{R}$ with $|\xi \mp \pi/2| < \pi$. This follows e.g. from [3, p. 273] by transformations, but may also be easily verified directly. Using (1.12) as definition, one immediately finds (1.11) by means of Watson's lemma. (See e.g. [11, p. 114] or [14, p. 40]).

From (1.12) we also obtain that the $H^{\pm}_{\kappa,\mu}$ are holomorphic with respect to (z,κ,μ) and satisfy

$$H^{\pm}_{\kappa,-\mu} = H^{\pm}_{\kappa,\mu} \tag{1.13}$$

In the very special situation of the DE (D_μ) one can determine the connection between the Floquet solutions $I_{\kappa,\pm\mu}$ and the asymptotic solutions $H^{\pm}_{\kappa,\mu}$ explicitly. Defining

$$I^{\pm}_{\kappa,\mu} := Q_{-\mu} H^{\pm}_{\kappa,\mu}, \tag{1.14}$$

we obtain by (1.8.1)

$$\frac{\sin 2\pi\mu}{\pi} H^{\pm}_{\kappa,\mu} = I^{\pm}_{\kappa,\mu} - I^{\pm}_{\kappa,-\mu} \tag{1.15}$$

and by (1.8.2) $I^{\pm}_{\kappa,\mu}\in \mathfrak{U}_{\mu}\cap \mathfrak{U}(D_{\mu})$. Using the connection formula between the Φ and Ψ functions (see e.g. [3, p. 257]) we finally find

$$I^{\pm}_{\kappa,\mu}=-(2e^{\mp \pi i/2})^{1/2\pm\kappa+\mu}\cdot\frac{1}{\Gamma(1/2\pm\kappa-\mu)}\cdot I_{\kappa,\mu}. \tag{1.16}$$

1.3. Series in terms of Floquet solutions of the DE (0.2). Throughout this section let $\nu\in\mathbb{C}$ with $2\nu\notin\mathbb{Z}$ and $\frac{1}{2}-\kappa+\nu\notin\mathbb{Z}$. The first condition implies $\mathfrak{U}_{\nu}\cap\mathfrak{U}_{-\nu}=\{0\}$ and both together $I^{\pm}_{\kappa,\mu}\neq 0$ for $\mu\in\pm\nu+\mathbb{Z}$.

For any two functions $y^{\pm}\in\mathfrak{U}_{\pm\nu}$ the product $y^{+}\cdot y^{-}$ is a holomorphic function on Ω. Hence we can define

$$\langle y^{+},y^{-}\rangle:=\langle y^{-},y^{+}\rangle:=\frac{1}{2\pi i}\oint_{|z|=\rho} y^{+}(z)y^{-}(z)\frac{dz}{z} \tag{1.17}$$

with an arbitrary $r<\rho<\infty$. Obviously, this expression is linear with respect to y^{+} and y^{-}.

Since the operator L_{κ} commutes with ϕ, the spaces $\mathfrak{U}_{\pm\nu}$ are invariant under L_{κ}. One easily verifies for $y^{\pm}\in\mathfrak{U}_{\pm\nu}$

$$\langle L_{\kappa}y^{+},y^{-}\rangle=\langle y^{+},L_{\kappa}y^{-}\rangle. \tag{1.18}$$

The results of the preceding section show that both eigenvalue problems

$$L_{\kappa}y^{\pm}=\lambda y^{\pm}\qquad(\lambda\in\mathbb{C};\ 0\neq y\in\mathfrak{U}_{\pm\nu})$$

possess the same eigenvalues, namely $\lambda=\mu^{2}$ with $\mu\in\nu+\mathbb{Z}$, and that the corresponding eigenfunctions are just $I^{+}_{\kappa,\mu}\in\mathfrak{U}_{\nu}$ and $I^{-}_{\kappa,-\mu}\in\mathfrak{U}_{-\nu}$. Using (1.18) as well as (1.16) and (1.9) then yields

$$\langle I^{+}_{\kappa,\mu+n},I^{-}_{\kappa,-\mu}\rangle=\delta_{n,0}\cdot\varepsilon_{\kappa,\mu}\qquad(\mu\in\nu+\mathbb{Z},\ n\in\mathbb{Z}) \tag{1.19}$$

where $\delta_{n,m}$ denotes the Kronecker symbol and

$$\varepsilon_{\kappa,\mu}:=e^{-i\pi(\kappa+\mu)}\cdot\frac{\sin\pi(1/2-\kappa+\mu)}{\pi}\cdot\frac{\sin 2\pi\mu}{\pi\mu}\neq 0\qquad(\mu\in\nu+\mathbb{Z}). \tag{1.20}$$

We now state a theorem which ensures the expansion of any function $y^{\pm}\in\mathfrak{U}_{\pm\nu}$ in terms of the eigenfunctions $I^{\pm}_{\kappa,\pm\mu}$ $(\mu\in\nu+\mathbb{Z})$. This theorem, which is fundamental for the following analysis, can be deduced easily from a corresponding theorem on expansions in terms of Whittaker functions found by F. W. Schäfke. (See e.g. [13, p. 177] or [14, p. 228].)

(1.21) THEOREM. *Each $y^{\pm}\in\mathfrak{U}_{\pm\nu}$ has an expansion*

$$y^{\pm}(z)=\sum_{\mu\in\pm\nu+\mathbb{Z}} c_{\mu}\cdot I^{\pm}_{\kappa,\mu}(z)\qquad(z\in\hat{\Omega}), \tag{1.21.1}$$

which converges absolutely and uniformly on compact subsets of $\hat{\Omega}$. The coefficients are uniquely determined by

$$c_{\mu}\cdot\varepsilon_{\kappa,\pm\mu}=\langle y^{\pm},I^{\mp}_{\kappa,-\mu}\rangle\qquad(\mu\in\pm\nu+\mathbb{Z}). \tag{1.21.2}$$

Moreover, the following asymptotic formula for the $I^{\pm}_{\kappa,\mu}$ holds. (See e.g. [13, p. 180] or [14, p. 245].)

(1.22) *Let for* $\mu \in \pm\nu + \mathbb{Z}$ *the functions* $g_{\kappa,\mu}: \mathbb{C} \to \mathbb{C}$ *be defined by*

$$I^{\pm}_{\kappa,\mu}(z) =: -(2e^{\mp i\pi/2})^{1/2\pm\kappa+\mu} \cdot \frac{z^{\mu}}{\Gamma(1+2\mu)\Gamma(1/2\pm\kappa-\mu)} g_{\kappa,\mu}(z).$$

Then

$$g_{\kappa,\pm\nu+n}(z) = 1 + O\left(\frac{1}{n}\right) \qquad (|n| \to \infty)$$

uniformly on compact subsets of $\mathbb{C}$.

Theorem (1.21) serves as a main tool to transform and study the DE (D) in suitable sequence spaces. For this purpose we make the following preliminary remarks.

Let $y^{\pm} \in \mathfrak{U}_{\pm\nu}$ and let $(c_{\mu})_{\mu \in \pm\nu+\mathbb{Z}}$ be the corresponding sequences of coefficients determined by (1.21.2). From (1.22) we obtain then

$$\lim_{n\to\infty} \left(|c_{\pm\nu+n}| \cdot (n!)^{-1}\right)^{1/n} = 0 \tag{1.23.1}$$

and

$$\limsup_{n\to\infty} \left(|c_{\pm\nu-n}| \cdot n!\right)^{1/n} \leqq \frac{r}{2}. \tag{1.23.2}$$

On the other hand, let $(c_{\mu})_{\mu \in \pm\nu+\mathbb{Z}}$ be sequences which behave like (1.23.1) and (1.23.2). Using (1.22) again, we find that the corresponding series (1.21.1) converge and define functions $y^{\pm} \in \mathfrak{U}_{\pm\nu}$. Moreover, the sequences $(c_{\mu})_{\mu \in \pm\nu+\mathbb{Z}}$ can be obtained back by (1.21.2). Introducing thus the vector spaces

$$\tilde{\mathfrak{U}}_{\pm\nu} := \left\{ c \in \mathbb{C}^{\pm\nu+\mathbb{Z}} \mid c \text{ behaves like (1.23.1) and (1.23.2)} \right\},$$

we have just seen that these are through (1.21.1), (1.21.2) isomorphic to $\mathfrak{U}_{\pm\nu}$, respectively.

For studying the DE (D) in $\tilde{\mathfrak{U}}_{\pm\nu}$, it becomes necessary to find the corresponding representations of the operators L_{κ} and G in $\tilde{\mathfrak{U}}_{\pm\nu}$. More generally, we consider an arbitrary differential operator of the form

$$Sy := \sum_{j=0}^{m} s_j \left(\frac{d}{dz}\right)^j y \qquad (y \in \mathscr{H}(\hat{\Omega})), \tag{1.24}$$

where $m \in \mathbb{N}$ and the coefficients s_j $(j = 0, \cdots, m)$ are holomorphic in Ω. Since S and ϕ commute, the spaces $\mathfrak{U}_{\pm\nu}$ are invariant under S. Hence, S has a representation in $\tilde{\mathfrak{U}}_{\pm\nu}$ which we denote by $\tilde{S}$. Letting $c = (c_{\mu})_{\mu \in \pm\nu+\mathbb{Z}} \in \tilde{\mathfrak{U}}_{\pm\nu}$ and $d = (d_{\mu})_{\mu \in \pm\nu+\mathbb{Z}} := \tilde{S}c \in \tilde{\mathfrak{U}}_{\pm\nu}$, we immediately obtain by (1.21)

$$d_{\mu} = \frac{1}{\varepsilon_{\kappa,\pm\mu}} \cdot \sum_{l\in\mathbb{Z}} \left\langle SI^{\pm}_{\kappa,\mu+l}, I^{\mp}_{\kappa,-\mu} \right\rangle \cdot c_{\mu+l} \qquad (\mu \in \pm\nu + \mathbb{Z}). \tag{1.25}$$

An important role in the following analysis plays the explicit knowledge of the representations of the special operators $1/z \cdot$ and d/dz in $\tilde{\mathfrak{U}}_{\pm\nu}$. These result from the following three-term recurrence relations

$$\begin{aligned} \frac{\mu}{z} I^{\pm}_{\kappa,\mu} &= \frac{i}{2}\left(\mp \frac{\mu \pm \kappa + 1/2}{\mu + 1/2} I^{\pm}_{\kappa,\mu+1} + \frac{2\kappa\mu}{\mu^2 - 1/4} I^{\pm}_{\kappa,\mu} \pm \frac{\mu \mp \kappa - 1/2}{\mu - 1/2} I^{\pm}_{\kappa,\mu-1} \right), \\ \frac{d}{dz} I^{\pm}_{\kappa,\mu} &= \frac{i}{2}\left(\pm \frac{\mu \pm \kappa + 1/2}{\mu + 1/2} I^{\pm}_{\kappa,\mu+1} + \frac{\kappa}{\mu^2 - 1/4} I^{\pm}_{\kappa,\mu} \pm \frac{\mu \mp \kappa - 1/2}{\mu - 1/2} I^{\pm}_{\kappa,\mu-1} \right), \end{aligned} \tag{1.26}$$

which can readily be obtained from [2, p. 82].

By the use of (1.26) we derive

(1.27) PROPOSITION. *Let S be an operator of the form* (1.24) *with coefficients s_j being holomorphic in $\Omega \cup \{\infty\}$. Then for $\mu \in \pm\nu + \mathbb{Z}$ and $n \in \mathbb{Z}$*

$$\frac{\Gamma(1/2 \pm \kappa - \mu - n)}{\Gamma(1/2 \pm \kappa - \mu)} \left\langle SI^{\pm}_{\kappa,\mu+n}, I^{\mp}_{\kappa,-\mu} \right\rangle = \frac{\Gamma(1/2 \pm \kappa - \mu)}{\Gamma(1/2 \pm \kappa - \mu - n)} \left\langle SI^{\mp}_{\kappa,-\mu-n}, I^{\pm}_{\kappa,\mu} \right\rangle.$$

The proof follows in several steps:

1) For $S = identity$ the formula is trivial by (1.19). For $S = 1/z \cdot$ and $S = d/dz$ it can directly be verified by using (1.26) and (1.19). In the first case both sides vanish for $|n| \geqq 1$ and in the latter for $|n| \geqq 2$.

2) Let now S_1 and S_2 be operators of the form (1.24) for which the formula (1.27) is valid. Obviously, it holds then for $S_1 + S_2$, too. A straightforward calculation using (1.25) shows that it also holds for $S_2 \circ S_1$.

3) According to 1) and 2) formula (1.27) is valid for operators of the form (1.24) where the s_j are polynomials in $1/z$. But then the general case immediately follows by a simple limiting process.

We finally prove

(1.28) PROPOSITION. *Let $0 < \tau < \rho < \infty$. Then for each $l \in \mathbb{Z}$ there exists a constant $\gamma_l = \gamma_l(\tau, \rho) > 0$ such that for all $k \in \mathbb{N}$ and all $n \in \mathbb{Z}$*

$$\left| \frac{1}{\varepsilon_{\kappa,\nu\pm n}} \left\langle z^{-k} I^{\pm}_{\kappa,\pm\nu+l}, I^{\mp}_{\kappa,\mp\nu-n} \right\rangle \right| \leqq \gamma_l \cdot \frac{\rho^{|n|}}{|n|!} (2\tau)^{-k}.$$

For $n \leqq N$ the inequality follows directly with (1.17) and (1.22) by estimating the integral at $|z| = 2\tau$. The case $n > N$ then can be reduced to the former case by means of (1.27). Here N has to be chosen large enough such that (1.27) is applicable.

1.4. Series in terms of asymptotic solutions of the DE (0.2). Throughout this section let $\nu, \kappa \in \mathbb{C}$ be arbitrary. We prove in the following

(1.29) THEOREM. *Let $(c_n) \in \mathbb{C}^{\mathbb{Z}}$ satisfy*

$$\limsup_{n \to \pm\infty} \left(|c_n| \cdot |n|! \right)^{1/|n|} \leqq \frac{r}{2}.$$

Then the series

$$H^{\pm}(z) := \sum_{n \in \mathbb{Z}} c_n \cdot H^{\pm}_{\kappa,\nu+n}(z) \qquad (z \in \hat{\Omega}) \tag{1.29.1}$$

define holomorphic functions in $\hat{\Omega}$ which possess the asymptotic expansions

$$H^{\pm}(z) \sim e^{\pm iz} \cdot z^{-1/2 \mp \kappa} \cdot \sum_{m=0}^{\infty} \frac{\tilde{c}^{\pm}_m}{m!} \cdot (\pm 2iz)^{-m} \tag{1.29.2}$$

as $|z| \to \infty$ within the sectors $|\arg z \mp \pi/2| < 3\pi/2$, where the coefficients $\tilde{c}^{\pm}_m$ are given by

$$\tilde{c}^{\pm}_m := \sum_{n \in \mathbb{Z}} c_n \cdot \left(\frac{1}{2} \pm \kappa + \nu + n \right)_m \cdot \left(\frac{1}{2} \pm \kappa - \nu - n \right)_m. \tag{1.29.3}$$

By the use of (1.15) and (1.22) it follows that the series (1.29.1) converge absolutely and uniformly on compact subsets of $\hat{\Omega}$ and hence define holomorphic functions in $\hat{\Omega}$. It remains to prove the asymptotic behavior. This may be done by representing $H^{\pm}$ as a Laplace integral and then applying Watson's lemma.

To simplify the following consideration let us assume that $c_n = 0$ for $n < N$, where $N \in \mathbb{N}$ with $\mathrm{Re}(\frac{1}{2} \pm \kappa + \nu + N) \geqq 1$. This means no restriction since by (1.13) one achieves at once $c_n = 0$ for $n < 0$ and the remaining finite series can be tackled directly by means of (1.11).

The representation (1.12) of $H^{\pm}_{\kappa,\mu}$ suggests that one study the series

$$f^{\pm}(\zeta) := \sum_{n=N}^{\infty} c_n \, {}_2F_1\left(\frac{1}{2} \pm \kappa + \nu + n, \, \frac{1}{2} \pm \kappa - \nu - n; \, 1; \, \zeta\right) \tag{1.30}$$

for $\zeta \in \mathbb{C} \setminus [1, +\infty[$. In view of this we prove

(1.31) LEMMA. *Let $a, b \in \mathbb{C}$ and $m \in \mathbb{N}$ with $\mathrm{Re}\, a \geqq 1$ and $m \geqq -\mathrm{Re}\, b$. Further, let $0 < \delta < 1$, $0 < \varphi < \pi/2$, $4 < \sigma < \infty$. Then there exist constants $0 < M < \infty$, $0 < \gamma < \infty$ and $1 < \rho < \infty$ such that for all $\zeta \in \mathbb{C}$ with $|\zeta| \leqq \delta$ or $\varphi \leqq \arg \zeta \leqq 2\pi - \varphi$ and for all $n \in \mathbb{N}$*

$$|{}_2F_1(a+n, \, b-n; \, 1; \, \zeta)| \leqq M \cdot \begin{cases} \sigma^n |\zeta|^{n+m} & (|\zeta| \geqq \rho), \\ \gamma^n & (|\zeta| \leqq \rho). \end{cases}$$

Proof. For $\zeta \in \mathbb{C} \setminus [1, +\infty[$ and $n \in \mathbb{N}$ the following integral representation is valid:

$${}_2F_1(a+n, \, b-n; \, 1; \, \zeta) = \frac{e^{i\pi(a+n)}}{2\pi i} \int_{\mathfrak{C}} t^{a-1} \cdot (1-t)^{-a} \cdot (1-t\zeta)^{-b} \cdot \left(\frac{t}{1-t}(1-t\zeta)\right)^n dt.$$

Here $\mathfrak{C}$ is a contour starting and ending at 0 and encircling 1 once in the positive sense which, moreover, is to be suitably chosen with respect to ζ. All powers are to be taken with their principal values near the initial point $t = 0$. (See e.g. [3, p. 60].)

A convenient choice for our purposes is given by the contour $\mathfrak{C}_\varepsilon$ (with $0 < \varepsilon < 1$) comprising the line segments $\overline{0\varepsilon}$ and $\overline{\varepsilon 0}$ together with the circle $|\zeta - 1| = 1 - \varepsilon$ taken in the positive sense. We intend to estimate the four factors of the integrand separately along $\mathfrak{C}_\varepsilon$. To this end let for the present $0 < \varepsilon < 1$ and $0 < \rho < \infty$ be arbitrary.

Then, for $t \in \mathfrak{C}_\varepsilon$

$$(\alpha) \qquad |t| \leqq 2 - \varepsilon, \qquad 1 - \varepsilon \leqq |1 - t| \leqq 1.$$

Furthermore, for $t \in \mathfrak{C}_\varepsilon$ and $\zeta \in \mathbb{C}$

$$(\beta 1) \qquad |1 - t\zeta| \leqq 2|\zeta| \qquad \text{if } |\zeta| \geqq \rho \geqq \frac{1}{\varepsilon},$$

$$(\beta 2) \qquad |1 - t\zeta| \leqq 1 + (2-\varepsilon)\rho \qquad \text{if } |\zeta| \leqq \rho,$$

as well as

$$(\gamma 1) \qquad |1 - t\zeta| \geqq \sin\varphi \qquad \text{if } 0 \leqq t \leqq \varepsilon, \quad \varphi \leqq \arg\zeta \leqq 2\pi - \varphi,$$

$$(\gamma 2) \qquad |1 - t\zeta| \geqq \rho\varepsilon - 1 \qquad \text{if } |t-1| \geqq 1 - \varepsilon, \quad |\zeta| \geqq \rho > \frac{1}{\varepsilon},$$

$$(\gamma 3) \qquad |1 - t\zeta| \geqq 1 - \delta(2-\varepsilon) \qquad \text{if } |\zeta| \leqq \delta,$$

$$(\gamma 4) \qquad |1 - t\zeta| \geqq \sin\frac{\varphi}{2} \qquad \text{if } \varphi \leqq \arg\zeta \leqq 2\pi - \varphi, \quad 1 - \varepsilon \leqq \sin\frac{\varphi}{2},$$

and

$$(\gamma 5) \qquad |\arg(1 - t\zeta)| \text{ bounded in each of the latter four cases.}$$

To prove the first inequality in (1.31), we choose $0<\varepsilon_1<1$ with $2(2-\varepsilon_1)/(1-\varepsilon_1)<\sigma$ and after that $1/\varepsilon_1<\rho<\infty$. Estimating the integral along $\mathfrak{C}_{\varepsilon_1}$ by the use of (α), $(\beta 1)$ $(\gamma 1)$, $(\gamma 2)$ and $(\gamma 5)$ then yields the desired result. To prove the second inequality in (1.31), we choose now $0<\varepsilon_2<1$ with $\delta\cdot(2-\varepsilon_2)<1$ and $1-\varepsilon_2\leqq\sin(\varphi/2)$. Estimating the integral along $\mathfrak{C}_{\varepsilon_2}$ by the use of (α), $(\beta 2)$, $(\gamma 3)$, $(\gamma 4)$ and $(\gamma 5)$ then yields the inequality for $|\zeta|\leqq\rho$. This completes the proof of the lemma.

We continue now in the proof of (1.29). Lemma (1.31) shows that the series (1.30) converge absolutely and uniformly on compact subsets of $\mathbb{C}\setminus[1,+\infty[$ and hence define holomorphic functions $f^{\pm}:\mathbb{C}\setminus[1,+\infty[\to\mathbb{C}$. One readily verifies the following power series representation at 0:

$$(*)\qquad f^{\pm}(\zeta)=\sum_{m=0}^{\infty}\frac{\tilde{c}_m^{\pm}}{m!}\frac{\zeta^m}{m!}\qquad(|\zeta|<1).$$

Using the first inequality in (1.31) again, one obtains, on the other hand, for arbitrary $0<\varphi<\pi/2$ and $r<\tilde{r}<\infty$ the inequality

$$(**)\qquad |f^{\pm}(\zeta)|\leqq\exp(2\tilde{r}|\zeta|),$$

being valid for all $\zeta\in\mathbb{C}$ with $\varphi\leqq\arg\zeta\leqq 2\pi-\varphi$ and $|\zeta|\geqq\rho$, where $1<\rho<\infty$ is to be chosen sufficiently large. Obviously, the right member in $(**)$ is also a bound for all partial sums of the series (1.30). This especially permits one to interchange summation and integration in the situation (1.29.1), (1.12) by means of the Lebesgue dominated convergence theorem and finally leads to the Laplace integral

$$(1.32)\qquad H^{\pm}(z)=e^{\pm iz}\cdot z^{1/2\mp\kappa}\cdot\int_0^{\infty\cdot e^{-i\xi}}e^{-zt}\cdot f^{\pm}\left(\pm\frac{t}{2i}\right)dt,$$

which is then valid for $z\in\hat{\Omega}$ with $\mathrm{Re}\,ze^{-i\xi}>r$ where $\xi\in\mathbb{R}$ with $|\xi\mp\pi/2|<\pi$.

Having $(*)$ and $(**)$ available, one can now immediately apply Watson's lemma onto the situation (1.32). This yields then the asymptotic expansions (1.29.2) as $|z|\to\infty$ within the half-planes $\mathrm{Re}\,ze^{-i\xi}>r$ where $\xi\in\mathbb{R}$ with $|\xi\mp\pi/2|<\pi$. To obtain the asymptotic expansion within the full sectors $|\arg z\mp\pi/2|<3\pi/2$, one has to choose an appropriate covering.

2. The global representation of the solutions of the DE (0.1) by series in terms of solutions of the DE (0.2). Throughout this chapter we assume that the DE (D) is in the "normal case", which means $2\nu\notin\mathbb{Z}$ for (one and hence for all) $\nu\in\Xi$. Then, for every $\nu\in\Xi$, we have $\frac{1}{2}-\kappa+\nu\notin\mathbb{Z}$ or $\frac{1}{2}+\kappa+\nu\notin\mathbb{Z}$ (since the sum of both quantities is $2\nu+1\notin\mathbb{Z}$). Considering $\nu\in\Xi\Leftrightarrow-\nu\in\Xi$, we may thus assume that a

$$\nu\in\Xi\text{ with }2\nu\notin\mathbb{Z}\quad\text{and}\quad\frac{1}{2}-\kappa+\nu\notin\mathbb{Z}$$

is chosen and fixed for the following. (This is done with regard to the unrestricted applicability of the results in §1.3). We then especially have $\Xi=(\nu+\mathbb{Z})\cup(-\nu+\mathbb{Z})$.

2.1. Transformation of the DE (0.1) into sequence spaces. According to (1.8.1), (1.8.2) each $y\in\mathfrak{N}(D)$ has a (unique) representation $y=y^{+}+y^{-}$ with $y^{\pm}\in\mathfrak{U}_{\pm\nu}\cap\mathfrak{N}(D)$. We shall thus consider in the following always (pairs of) functions $y^{\pm}\in\mathfrak{U}_{\pm\nu}$.

Let $y^{\pm}\in\mathfrak{U}_{\pm\nu}\cap\mathfrak{N}(D)$. Then, according to §1.3, the corresponding sequences $(c_\mu)_{\mu\in\pm\nu+\mathbb{Z}}\in\tilde{\mathfrak{U}}_{\pm\nu}$ defined by (1.21.2) are solutions of the equation

$$\tilde{L}_\kappa c+\tilde{G}c=0,$$

which can also be written in the more explicit form

$$\mu^2 \cdot c_\mu + \frac{1}{\varepsilon_{\kappa,\pm\mu}} \cdot \sum_{l\in\mathbb{Z}} \left\langle G I^{\pm}_{\kappa,\mu+l}, I^{\mp}_{\kappa,-\mu} \right\rangle \cdot c_{\mu+l} = 0, \qquad (\mu \in \pm\nu + \mathbb{Z}). \tag{2.1}$$

This follows immediately from (1.25) with $S = L_\kappa$ and $S = G$; the off-diagonal terms of $\tilde{L}_\kappa$ vanish due to $L_\kappa I^{\pm}_{\kappa,\mu} = \mu^2 I^{\pm}_{\kappa,\mu}$ and (1.19).

In the following, we also need the converse statement: Let the sequences $(c_\mu)_{\mu\in\pm\nu+\mathbb{Z}}$ be solutions of (2.1) belonging to $\tilde{\mathfrak{U}}_{\pm\nu}$, which means by (1.23) that they have the asymptotic behavior (1.23.1) and (1.23.2). Then, according to §1.3, the corresponding functions $y^{\pm} \in \mathfrak{U}_{\pm\nu}$ defined by (1.21.1) are solutions of (D), so $y^{\pm} \in \mathfrak{U}_{\pm\nu} \cap \mathfrak{N}(D)$.

The main aim of the following analysis is to improve the asymptotic behavior (1.23.1) for the solutions of the equation (2.1). This means to show that the solutions of (2.1) belong to other specific sequence spaces. We thus have to put the equation into a suitable form such that it can also be treated in the other sequence spaces in question.

For this purpose we define, for the moment only for sequences $c = (c_\mu)_{\mu\in\pm\nu+\mathbb{Z}} \in \tilde{\mathfrak{U}}_{\pm\nu}$,

$$(Tc)_\mu := \frac{1}{\mu} c_\mu, \tag{2.2}$$

$$(Ac)_\mu := \frac{1}{\varepsilon_{\kappa,\pm\mu}} \cdot \sum_{l=-1}^{1} \left\langle \frac{1}{z} \cdot I^{\pm}_{\kappa,\mu+l},\ I^{\mp}_{\kappa,-\mu} \right\rangle \cdot c_{\mu+l}, \tag{2.3}$$

$$(Bc)_\mu := \frac{1}{\mu \cdot \varepsilon_{\kappa,\pm\mu}} \cdot \sum_{l=-1}^{1} \left\langle \frac{d}{dz} \cdot I^{\pm}_{\kappa,\mu+l},\ I^{\mp}_{\kappa,-\mu} \right\rangle \cdot c_{\mu+l}, \tag{2.4}$$

$$(G_m c)_\mu := \frac{1}{\varepsilon_{\kappa,\pm\mu}} \cdot \sum_{l\in\mathbb{Z}} \left\langle g_m I^{\pm}_{\kappa,\mu+l},\ I^{\mp}_{\kappa,-\mu} \right\rangle \cdot c_{\mu+l}. \tag{2.5}$$

$(\mu \in \pm\nu + \mathbb{Z})$

Obviously, the sequences Tc, Ac, Bc and $G_m c$ belong to $\tilde{\mathfrak{U}}_{\pm\nu}$ and, hence, T, A, B and G_m are operators in $\tilde{\mathfrak{U}}_{\pm\nu}$. With the notation of §1.3 we especially have

$$T^2 = \tilde{L}_k^{-1}, \quad A = \left(\widetilde{\frac{1}{z}\cdot}\right), \quad B = T \circ \left(\widetilde{\frac{d}{dz}}\right), \quad G_m = \left(\widetilde{g_m \cdot}\right).$$

Thus, (2.1) is equivalent to

$$c + T \circ (B \circ G_1 + T \circ G_2) c = 0. \tag{2.6}$$

2.2. Solution of the DE (0.1) in sequence spaces. Let $0 < \rho < \infty$. By means of

$$\omega_1(n) := \frac{\rho^{|n|}}{|n|!}, \qquad \omega_2(n) := \frac{|n|!}{\rho^{|n|}} \qquad (n \in \mathbb{Z})$$

we define for $j = 1, 2$ the sequence spaces

$$\mathfrak{M}_j^{\pm} := \left\{ (c_\mu) \in \mathbb{C}^{\pm\nu+\mathbb{Z}} \,\middle|\, \exists \gamma > 0 \text{ s.t. } |c_{\pm\nu+n}| \leqq \gamma \cdot \omega_j(n),\ (n \in \mathbb{Z}) \right\}$$

and for $c = (c_\mu) \in \mathfrak{M}_j^{\pm}$ respectively the norms

$$\|c\|_j := \min\left\{ \gamma \geqq 0 \,\middle|\, |c_{\pm\nu+n}| \leqq \gamma \cdot \omega_j(n),\ (n \in \mathbb{Z}) \right\}.$$

Obviously, the $\mathfrak{M}_j^{\pm}$ (with corresponding norm $\|\cdot\|_j$) are Banach spaces and

$$\mathfrak{M}_1^{\pm} \hookrightarrow \mathfrak{M}_2^{\pm} \quad \text{continuously imbedded.}$$

We intend to study the equation (2.6) in the spaces $\mathfrak{M}_j^{\pm}$, $(j=1,2)$. We thus show at first

(2.7) PROPOSITION. *The formulas* (2.2), (2.3) *and* (2.4) *define continuous operators* T, A *and* B *in each of the spaces* $\mathfrak{M}_j^{\pm}$ $(j=1,2)$.

Proof. The case (2.2) is obvious. Since the cases (2.3) and (2.4) are quite analogous, we restrict ourselves to the case (2.3). Let $c=(c_\mu)\in\mathfrak{M}_j^{\pm}$. Using (1.26) we obtain for $\mu\in\pm\nu+\mathbb{Z}$

$$(Ac)_\mu = \pm\frac{i}{2}\left(1\mp\frac{\kappa}{\mu+1/2}\right)\frac{c_{\mu+1}}{\mu+1}+\frac{i\kappa}{\mu^2-1/4}c_\mu\mp\frac{i}{2}\left(1\pm\frac{\kappa}{\mu-1/2}\right)\cdot\frac{c_{\mu-1}}{\mu-1}.$$

This immediately yields the inequality

$$|(Ac)_{\pm\nu+n}|\leqq\frac{1}{2\rho}\cdot\sigma_n\cdot\omega_j(n)\cdot\|c\|_j \qquad (n\in\mathbb{Z}), \tag{2.8.1}$$

where the σ_n are independent of c and

$$\mathbb{R}^+\ni\sigma_n\to 1 \qquad (|n|\to\infty). \tag{2.8.2}$$

Especially, for $c\in\mathfrak{M}_j^{\pm}$

$$Ac\in\mathfrak{M}_j^{\pm} \quad\text{and}\quad \|Ac\|_j\leqq\frac{1}{2\rho}\cdot\sup_{n\in\mathbb{Z}}\sigma_n\cdot\|c\|_j.$$

The case (2.5) is more difficult. We shall essentially show that like the g_m are power series in $1/z$ the corresponding G_m are power series in A.

We consider for $N\in\mathbb{N}$ the projection P_N which we define for $c=(c_\mu)\in\mathfrak{M}_j^{\pm}$ by

$$(P_Nc)_{\pm\nu+n} := \begin{cases} c_{\pm\nu+n} & (|n|>N),\\ 0 & (|n|\leqq N).\end{cases}$$

Obviously, the P_N are continuous operators in $\mathfrak{M}_j^{\pm}$ satisfying

$$\|P_N\|_j\leqq 1 \quad\text{and}\quad \|1-P_N\|_j\leqq 1.$$

Here, $\|\cdot\|_j$ denotes the corresponding operator norm. From (2.8.1) and (2.8.2) we immediately obtain

(2.9) COROLLARY. *To each* $0<\tau<\rho$ *there exists a* $N\in\mathbb{N}$ *with*

$$\|A\circ P_N\|_j\leqq\frac{1}{2\tau}.$$

We can prove now

(2.10) PROPOSITION. *In the case* $r/2<\rho<\infty$ *the formula* (2.5) *defines continuous operators* G_1 *and* G_2 *in each of the spaces* $\mathfrak{M}_j^{\pm}$ $(j=1,2)$.

Proof. Let $r/2<\tau<\sigma<\rho$ be fixed and $N\in\mathbb{N}$ be chosen to τ according to (2.9).

(i) Using (1.25), one readily finds for $k\in\mathbb{N}$ and $c=(c_\mu)\in\mathfrak{M}_j^{\pm}$

$$(A^kc)_\mu=\frac{1}{\varepsilon_{\kappa,\pm\mu}}\cdot\sum_{|l|\leqq k}\left\langle z^{-k}I^{\pm}_{\kappa,\mu+l},\ I^{\mp}_{\kappa,-\mu}\right\rangle\cdot c_{\mu+l} \qquad (\mu\in\pm\nu+\mathbb{Z}). \tag{2.11}$$

Estimating with (1.28), this yields for $c=(c_\mu)\in\operatorname{range}(1-P_N)$

$$|(A^k c)_{\pm\nu+n}|\leqq\Big(\sum_{|l|\leqq N}\gamma_l\cdot|c_{\pm\nu+l}|\Big)\cdot\omega_1(n)\cdot(2\sigma)^{-k}\qquad(n\in\mathbb{Z}),$$

where $\gamma_l=\gamma_l(\sigma,\rho)$. Thus there is a constant $1\leqq\tilde\gamma<\infty$ with

$$\|A^k\circ(1-P_N)\|_j\leqq\tilde\gamma\cdot(2\sigma)^{-k}. \tag{2.12}$$

By induction now immediately follows

$$\|A^k\|_j\leqq\tilde\gamma\cdot(2\tau)^{-k}\cdot\prod_{j=1}^{k}\Big(1+\Big(\frac{\tau}{\sigma}\Big)^j\Big)\leqq\gamma\cdot(2\tau)^{-k}\qquad(k\in\mathbb{N}), \tag{2.13}$$

where $\gamma:=\tilde\gamma\cdot\exp(\tau/(\sigma-\tau))$. For $k=0$ this is true on account of $\tilde\gamma\geqq1$. For the step "$k\to k+1$" we estimate by the use of (2.12) and (2.9)

$$\begin{aligned}\|A^{k+1}\|_j&\leqq\|A^k\|_j\|A\circ P_N\|_j+\|A^{k+1}\circ(1-P_N)\|_j\\&\leqq\tilde\gamma\cdot(2\tau)^{-k}\cdot\prod_{j=1}^{k}\Big(1+\Big(\frac{\tau}{\sigma}\Big)^j\Big)\cdot\frac{1}{2\tau}+\tilde\gamma\cdot(2\sigma)^{-k-1}.\end{aligned}$$

(ii) Expanding the g_m into power series

$$g_m(z)=:\sum_{k=0}^{\infty}g_k^m\cdot z^{-k}\qquad(r<|z|\leqq\infty),$$

we obtain

$$\sum_{k=0}^{\infty}|g_k^m|\cdot(2\tau)^{-k}=:\gamma_m<\infty.$$

By the use of (1.25), (2.11) and (2.13) then follow for $c=(c_\mu)\in\mathfrak{M}_j^{\pm}$ and $\mu=\pm\nu+n$

$$\begin{aligned}|(G_m c)_\mu|&\leqq\frac{1}{|\varepsilon_{\kappa,\pm\mu}|}\cdot\sum_{l\in\mathbb{Z}}\sum_{k\geqq|l|}|g_k^m|\Big|\big\langle z^{-k}I^{\pm}_{\kappa,\mu+l},\,I^{\mp}_{\kappa,-\mu}\big\rangle\Big|\cdot|c_{\mu+l}|\\&\leqq\sum_{k=0}^{\infty}|g_k^m|\cdot\frac{1}{|\varepsilon_{\kappa,\pm\mu}|}\cdot\sum_{|l|\leqq k}\Big|\big\langle z^{-k}I^{\pm}_{\kappa,\mu+l},\,I^{\mp}_{\kappa,-\mu}\big\rangle\Big|\cdot|c_{\mu+l}|\\&\leqq\Big(\sum_{k=0}^{\infty}|g_k^m|\cdot\|A^k\|_j\Big)\cdot\|c\|_j\cdot\omega_j(n)\leqq\gamma\cdot\gamma_m\cdot\|c\|_j\cdot\omega_j(n).\end{aligned}$$

Thus, for $c\in\mathfrak{M}_j^{\pm}$,

$$G_m c\in\mathfrak{M}_j^{\pm}\quad\text{and}\quad\|G_m c\|_j\leqq\gamma\cdot\gamma_m\cdot\|c\|_j.$$

According to (2.7) and (2.10) we can consider equation (2.6) now in $\mathfrak{M}_1^{\pm}$ as well as in $\mathfrak{M}_2^{\pm}$, provided that $r/2<\rho$. We prove

(2.14) PROPOSITION. *Let $r/2<\rho<\infty$. Then each solution of* (2.6) *belonging to $\mathfrak{M}_2^{\pm}$ already belongs to $\mathfrak{M}_1^{\pm}$.*

Proof. By the definition of T immediately follows $\|P_N\circ T\|_j\to0$ for $N\to\infty$. Hence we can choose $N\in\mathbb{N}$ such that $\|P_N\circ T\|_j\cdot\|B\circ G_1+T\circ G_2\|_j<1$. Applying P_N to (2.6) and adding $(1-P_N)c$ on both sides gives

$$(1-P_N)c=\big(1+(P_N\circ T)\circ(B\circ G_1+T\circ G_2)\big)c. \tag{$*$}$$

Since the left side of (*) obviously belongs to $\mathfrak{M}_1^{\pm}$ and the operator on the right side of (*) is, by the above choice of N, invertible in $\mathfrak{M}_1^{\pm}$ as well as in $\mathfrak{M}_2^{\pm}$, c necessarily belongs to $\mathfrak{M}_1^{\pm}$.

Finally, we apply the result of (2.14) to our special situation where we have a solution c of (2.6) belonging to $\tilde{\mathfrak{U}}_{\pm\nu}$. Choosing an arbitrary $(r/2)<\rho<\infty$, we obviously have $c\in\tilde{\mathfrak{U}}_{\pm\nu}\subset\mathfrak{M}_2^{\pm}$ and hence by (2.14) obtain $c\in\mathfrak{M}_1^{\pm}$. We summarize the essential facts in the following

(2.15) THEOREM. *Let* $y^{\pm}\in\mathfrak{U}_{\pm\nu}\cap\mathfrak{N}(D)$ *and let* $(c_\mu)_{\mu\in\pm\nu+\mathbb{Z}}$ *be the corresponding coefficients defined by* (1.21.2). *Then*

$$\limsup_{|n|\to\infty}\left(|c_{\pm\nu+n}|\cdot|n|!\right)^{1/n}\leqq\frac{r}{2}. \tag{2.15.1}$$

2.3. Floquet and asymptotic solutions of the DE (0.1). Let $0\neq y^{\pm}\in\mathfrak{U}_{\pm\nu}\cap\mathfrak{N}(D)$. According to (1.21) we have

$$y^{\pm}(z)=\sum_{\mu\in\pm\nu+\mathbb{Z}}c_\mu I^{\pm}_{\kappa,\mu}(z)\qquad(z\in\hat{\Omega}) \tag{2.16}$$

where the coefficients $(c_\mu)_{\mu\in\pm\nu+\mathbb{Z}}$ are given by (1.21.2). As shown in the last section, these coefficients possess the asymptotic behavior (2.15.1). We can thus define

$$\hat{y}^{\pm}(z):=\sum_{\mu\in\pm\nu+\mathbb{Z}}c_{-\mu}I^{\mp}_{\kappa,\mu}(z)=\sum_{\mu\in\pm\nu+\mathbb{Z}}\hat{c}_\mu I^{\pm}_{\kappa,\mu}(z)\qquad(z\in\hat{\Omega}) \tag{2.17}$$

where—by (1.16)—

$$\hat{c}_\mu:=c_{-\mu}\cdot 4^{\mp\kappa}\cdot\left(e^{\pm i\pi/2}\right)^{1+2\mu}\cdot\frac{\Gamma(1/2\pm\kappa-\mu)}{\Gamma(1/2\mp\kappa-\mu)}\qquad(\mu\in\pm\nu+\mathbb{Z}). \tag{2.17.1}$$

Obviously, $\hat{y}^{\pm}\in\mathfrak{U}_{\pm\nu}$. Multiplying (2.1) with

$$4^{\pm\kappa}\cdot\left(e^{\mp i\pi/2}\right)^{1-2\mu}\cdot\frac{\Gamma(1/2\mp\kappa+\mu)}{\Gamma(1/2\pm\kappa+\mu)}$$

and using (1.27) with $S=G$ one readily finds that $(\hat{c}_\mu)_{\mu\in\pm\nu+\mathbb{Z}}$ is a solution of (2.1) in $\tilde{\mathfrak{U}}_{\pm\nu}$. Hence, according to §2.1, $\hat{y}^{\pm}\in\mathfrak{U}_{\pm\nu}\cap\mathfrak{N}(D)$.

Since $y^{\pm}\neq 0$, there exist (unique) constants $\xi^{\pm}\in\mathbb{C}$ such that

$$\hat{y}^{\pm}(z)=\pm\frac{e^{\pi i\nu}}{2\sin\pi(1/2+\kappa-\nu)}\cdot\xi^{\pm}\cdot y^{\pm}(z)\qquad(z\in\hat{\Omega}). \tag{2.18}$$

This immediately implies

$$\xi^{\pm}\cdot c_\mu=\frac{2\pi i\cdot 4^{\mp\kappa}}{\Gamma(1/2\mp\kappa+\mu)\Gamma(1/2\mp\kappa+\mu)}\cdot c_{-\mu}\qquad(\mu\in\pm\nu+\mathbb{Z}), \tag{2.19}$$

which obviously is a connection formula between the coefficients of y^{+} and y^{-}. Using (2.19) twice gives

$$\xi^{+}\cdot\xi^{-}=-4\sin\pi\left(\frac{1}{2}+\kappa-\nu\right)\sin\pi\left(\frac{1}{2}+\kappa+\nu\right)=-2(\cos(2\pi\nu)+\cos(2\pi\kappa)). \tag{2.20}$$

A short calculation then yields

$$1+\frac{e^{2\pi i\nu}\cdot\xi^{+}\cdot\xi^{-}}{4\sin^2(1/2+\kappa-\nu)}=\frac{e^{\pi i(\kappa+\nu-1/2)}\cdot\sin 2\nu\pi}{\sin\pi(1/2+\kappa-\nu)}\neq 0. \tag{*}$$

We define now for $z \in \hat{\Omega}$

$$\frac{\sin 2\pi\nu}{\pi}(H^+(z), H^-(z)) := (y^+(z), y^-(z)) \cdot \begin{pmatrix} 1 & \frac{e^{\pi i\nu}\cdot\xi^+}{2\sin\pi(1/2+\kappa-\nu)} \\ \frac{e^{\pi i\nu}\cdot\xi^-}{2\sin\pi(1/2+\kappa-\nu)} & -1 \end{pmatrix}. \tag{2.21}$$

Since y^+ and y^- constitute a fundamental set of solutions of (D), also H^+ and H^- do so on account of $(*)$. By the use of (2.18), (2.17) and (1.15) we obtain the representations

$$H^{\pm}(z) = \sum_{\mu \in \pm\nu + \mathbb{Z}} c_\mu \cdot H^{\pm}_{\kappa,\mu}(z) \qquad (z \in \hat{\Omega}). \tag{2.22}$$

(1.29) then shows that the $H^{\pm}$ possess the asymptotic expansions

$$H^{\pm}(z) \sim e^{\pm iz}\cdot z^{\mp\kappa - 1/2}\cdot \sum_{m=0}^{\infty} \frac{\tilde{c}_m^{\pm}}{m!}(\pm 2iz)^{-m} \tag{2.23}$$

as $|z| \to \infty$ within the sectors $|\arg z \mp \pi/2| < 3\pi/2$, where the coefficients are given by

$$\tilde{c}_m^{\pm} := \sum_{\mu \in \pm\nu + \mathbb{Z}} c_\mu \cdot \left(\frac{1}{2} \pm \kappa + \mu\right)_m \cdot \left(\frac{1}{2} \pm \kappa - \mu\right)_m. \tag{2.23.1}$$

Inserting the asymptotic series (2.23) into the Wronskian we arrive after a short calculation at

$$[H^+, H^-] = -2i \cdot \tilde{c}_0^+ \cdot \tilde{c}_0^-.$$

Since H^+ and H^- constitute a fundamental system, we have $[H^+, H^-] \neq 0$ and thus obtain

$$\tilde{c}_0^{\pm} = \sum_{\mu \in \pm\nu + \mathbb{Z}} c_\mu \neq 0.$$

This makes it possible to normalize the solutions $y^{\pm}$ and $H^{\pm}$. A quite natural choice is $\tilde{c}_0^{\pm} = 1$. By this, $y^{\pm}$ and $H^{\pm}$ become uniquely determined. We summarize the results concerning the $H^{\pm}$.

(2.24) THEOREM. *There is a unique fundamental set of solutions H^+ and H^- of the DE (D) having a representation of the form* (2.22), *where the coefficients $(c_\mu)_{\mu \in \Xi}$ satisfy* (2.15.1) *and*

$$\sum_{\mu \in \pm\nu + \mathbb{Z}} c_\mu = 1. \tag{2.24.1}$$

The $H^{\pm}$ possess the asymptotic expansions (2.23), (2.23.1). *Furthermore, the coefficients of H^+ and H^- satisfy the connection formula* (2.19), *where*

$$\xi^{\pm} = \sum_{\mu \in \mp\nu + \mathbb{Z}} \frac{2\pi i \cdot 4^{\mp\kappa}}{\Gamma(1/2 \mp \kappa + \mu)\Gamma(1/2 \mp \kappa - \mu)} c_\mu. \tag{2.24.2}$$

The formula (2.24.2) follows by summing up (2.19) and using (2.24.1).

Introducing the *monodromy matrix* U of the fundamental set of solutions H^+ and H^- through

$$(\phi H^+(z),\phi H^-(z)) =: (H^+(z),H^-(z))\circ U \qquad (z\in\hat{\Omega}), \tag{2.25}$$

a short calculation using (2.21) and (*) yields

$$U=\begin{pmatrix} -e^{-2\pi i\kappa} & -e^{-\pi i\kappa}\cdot\xi^+ \\ -e^{-\pi i\kappa}\cdot\xi^- & 2\cos 2\pi\nu + e^{-2\pi i\kappa}\end{pmatrix}. \tag{2.25.1}$$

The formulas (2.25.1) and (2.20) show the actual meaning and the fundamental importance of the quantities $\xi^{\pm}$.

We shall now call attention to the special case where H^+ or H^- itself is a Floquet solution. A preliminary result which is closely related to this question is the following:

$$\Xi=\left(\frac{1}{2}+\kappa+\mathbb{Z}\right)\cup\left(\frac{1}{2}-\kappa+\mathbb{Z}\right) \quad \text{if and only if } \xi^+\cdot\xi^-=0. \tag{2.26}$$

This is obvious by (2.20). A more detailed information is contained in

(2.27) THEOREM. *A necessary and sufficient condition for $H^{\pm}$ to be a Floquet solution is $\xi^{\mp}=0$. In this case*

$$c_{\mp\kappa-1/2+n}=0 \qquad (\mathbb{Z}\ni n\geqq 1)$$

and hence

$$\begin{aligned} H^{\pm}(z) &= \sum_{n=0}^{\infty} c_{\mp\kappa-1/2-n}\cdot H^{\pm}_{\kappa,\pm\kappa+1/2+n}(z) \\ &= \frac{\pm\pi}{\sin 2\pi\kappa}\sum_{n=0}^{\infty} c_{\mp\kappa-1/2-n} I^{\pm}_{\kappa,\mp\kappa-1/2-n}(z). \end{aligned}$$

Moreover, the corresponding asymptotic expansion actually converges in $\hat{\Omega}$ and represents $H^{\pm}$.

The last statement follows from the fact that in this case $z^{1/2\pm\kappa}\cdot e^{\mp iz}\cdot H^{\pm}(z)$ is holomorphic in $\Omega\cup\{\infty\}$. The other statements follow directly by (2.21), (2.19) and (1.15), (1.16).

The following remark shall clarify the actual dependence of our functions on the previous choice of ν. Obviously, all definitions and calculations only depend on $\nu\in\Xi$ modulo 1, but some of them essentially use the assumptions $\frac{1}{2}-\kappa+\nu\notin\mathbb{Z}$. Regarding especially the solutions $H^{\pm}$, one easily sees that these are uniquely determined by their asymptotic behavior and hence are independent of ν. Then, by (2.25), (2.25.1), also the $\xi^{\pm}$ are independent of ν.

Since the $H^{\pm}$ are independent of the previous choice of ν, we can give an appropriate definition of Floquet solutions depending on the $H^{\pm}$, which is in full accordance with the definition (1.14).

For $\mu\in\Xi$ let

$$I^{\pm}_{\mu} := Q_{-\mu}H^{\pm}. \tag{2.28}$$

By (1.8.1), (1.8.2) then follows

$$\frac{\sin 2\pi\mu}{\pi}\cdot H^{\pm}=I_{\mu}^{\pm}-I_{-\mu}^{\pm},\qquad I_{\mu}^{\pm}\in\mathfrak{U}_{\pm\mu}\cap\mathfrak{N}(D). \tag{2.28.1}$$

Especially, for $\mu\in\Xi$ with $\frac{1}{2}-\kappa+\mu\notin\mathbb{Z}$, we obtain from (2.28.1), (2.21) and (2.18)

$$I_{\pm\mu}^{\pm}=y^{\pm}\neq 0;\ I_{\mp\mu}^{\pm}=\tilde{y}^{\mp}\qquad(\neq 0\Leftrightarrow\xi^{\mp}\neq 0). \tag{2.28.2}$$

We close our considerations with a few *remarks* concerning the practicality of our representations and their relation to other recent results on the subject.

Our series representations of the solutions of the DE (D) depend at first on the characteristic exponents $\nu\in\Xi$ and secondly on (essentially) one set of coefficients $(c_{\mu})_{\mu\in\pm\nu+\mathbb{Z}}$.

The characteristic exponents are determined by (2.20) through the quantities $\xi^{\pm}$. This fact—but even more their fundamental role as coefficients of the monodromy matrix U—makes it desirable having a method to compute the $\xi^{\pm}$. Unfortunately, (2.24.2) is not accessible for a direct computation because the $(c_{\mu})_{\mu\in\pm\nu+\mathbb{Z}}$ are not available. But (2.24.2) together with (2.23.1) yields the following limit-formula for the $\xi^{\pm}$

$$\lim_{n\to\infty}\frac{\tilde{c}_n^{\pm}}{\Gamma(n\pm 2\kappa)\cdot n!}=\frac{1}{2\pi i}\cdot 4^{\mp\kappa}\cdot\xi^{\mp}. \tag{2.29}$$

Since the coefficients $\tilde{c}_n^{\pm}$ of the asymptotic expansion (2.23) can be calculated recursively in a well-known manner, this formula provides an explicit method to determine approximately the $\xi^{\pm}$ (and thus the characteristic exponents $\nu\in\Xi$ and the monodromy matrix U). (2.29) is exactly the formula obtained by Jurkat–Lutz–Peyerimhoff in [6], resp. [7], [8] for calculating the so-called *Birkhoff invariants* of the DE (D), which are essentially ξ^{+} and ξ^{-}. Another method to determine the $\xi^{\pm}$ is to be found in Hinton [4].

Once knowing the characteristic exponents $\nu\in\Xi$, the coefficients $(c_{\mu})_{\mu\in\pm\nu+\mathbb{Z}}$ of our expansions are then uniquely determined by equation (2.1) and conditions (1.23.1), (1.23.2), (2.24.1). Unfortunately, equation (2.1) is, in general, not recursively solvable. The situation is nearly the same as for calculating the coefficients of the Laurent expansion of the Floquet solutions $y^{\pm}$. But this is not at all astonishing: Knowing the power series expansion of the $I_{\kappa,\mu}^{\pm}$ explicitly, one can express the Laurent coefficients of the $y^{\pm}$ through our coefficients c_{μ} and vice versa. Nevertheless, if the coefficients g_1 and g_2 of the DE (D) are polynomials in $1/z$, equation (2.1) becomes a difference-equation, which can be solved recursively with respect to the conditions (1.23.1), (1.23.2). Herein are included many types of confluent Fuchsian DEs which are of interest in the theory of higher special functions.

REFERENCES

[1] G. D. Birkhoff, *On a simple type of irregular singular point*, Trans. Amer. Math. Soc., 14 (1913), pp. 462–476.

[2] H. Buchholz, *The Confluent-Hypergeometric Function*, Springer-Verlag, Berlin, Heidelberg, New York, 1969.

[3] Erdelyi Magnus, et al., *Higher Transcendental Functions*, Bateman manuscript project, vol. 1, McGraw-Hill, New York, 1953.

[4] F. L. Hinton, *Stokes multipliers for a class of ordinary differential equations*, J. Math. Phys., 20 (1979), pp. 2036–2046.

[5] J. HORN, *Über die asymptotische Darstellung der Integrale linearer Differentialgleichungen*, J. Reine Angew. Math., 133 (1908), pp. 19–67.

[6] W. JURKAT, D. A. LUTZ AND A. PEYERIMHOFF, *Invariants and canonical forms for meromorphic second order differential equations*, Proc. 2nd Scheveningen Conference on Differential Equations, Math. Studies no. 21, North-Holland, Amsterdam, pp. 181–187.

[7] _____, *Birkhoff invariants and effective calculations for meromorphic linear differential equations*, I, J. Math. Anal., 53 (1976), pp. 438–470.

[8] _____, *Birkhoff invariants and effective calculations for meromorphic linear differential equations*, II, Houston J. Math., 2 (1976), pp. 207–238.

[9] J. MEIXNER AND F. W. SCHÄFKE, *Mathieusche Funktionen und Sphäroidfunktionen*, Springer-Verlag, Berlin, Göttingen, Heidelberg, 1954.

[10] J. MEIXNER, F. W. SCHÄFKE AND G. WOLF, *Mathieu Functions and Spheroidal Functions and Their Mathematical Foundations*, Lecture Notes in Mathematics 837, Springer-Verlag, Berlin, Göttinger, Heidelberg, 1980.

[11] F. W. J. OLVER, *Asymptotics and Special Functions*, Academic Press, New York and London, 1974.

[12] _____, *On the asymptotic solutions of second order differential equations having an irregular singularity of rank one, with an application to Whittaker functions*, SIAM J. Numer. Anal., 2 (1965), pp. 225–243.

[13] F. W. SCHÄFKE, *Reihenentwicklungen analytischer Funktionen nach Biorthogonalsystemen spezieller Funktionen*, II, Math. Z., 75 (1961), pp. 154–191.

[14] _____, *Einführung in die Theorie der Speziellen Funktionen der Mathematischen Physik*, Springer-Verlag, Berlin, Göttingen, Heidelberg, 1963.

[15] _____, *Zur Theorie der Neumannschen und Kapteynschen Reihen*, Arch. d. Math., 34 (1980), pp. 132–139.

[16] D. SCHMIDT, *Die Lösung der linearen Differentialgleichung 2. Ordnung um zwei einfache Singularitäten durch Reihen nach hypergeometrischen Funktionen*, J. Reine Angew. Math., 309 (1979), pp. 127–148.

SIAM J. MATH. ANAL.
Vol. 17, No. 5, September 1986

005

DISCONJUGACY AND COMPARISON THEOREMS FOR SECOND-ORDER LINEAR SYSTEMS*

W. J. KIM†

Abstract. Sufficient conditions for disconjugacy and disfocality are obtained for the second-order system $y''+Ay=0$, where A is an $n\times n$ matrix. Also proved are comparison theorems for disfocality through a generalization of the Riccati equation method originally used by Hille (Trans. Amer. Math. Soc., 64 (1948), pp. 234–252) for the case $n=1$.

1. Introduction. The second-order systems to be studied in this paper are of the form

$$y''+Ay=0, \qquad A=(a_{ij})_{i,j=1}^{n}, \tag{1}$$

where $A=A(x)$ is an $n\times n$ matrix with real elements which are continuous on an x-interval I, and y is an n-dimensional column vector. The system (1) is said to be *disconjugate* on the interval I if, for every pair of points a, $b\in I$, $a<b$, the only solution y satisfying the two-point boundary condition $y(a)=y(b)=0$ is the trivial one. On the other hand, if the only solution of (1) satisfying the condition $y(a)=y'(b)=0$, $a,b\in I$, $a<b$, is the trivial solution, then (1) is said to be *right-disfocal* on I. Similarly, we shall say that (1) is *left-disfocal* on I if no nontrivial solution satisfies the condition $y'(a)=y(b)=0$, $a,b\in I$, $a<b$.

The concepts of disconjugacy and disfocality defined for the second-order system (1) are closely related to the concepts of suborthogonality and nonoscillation that Nehari [13] defined for the first-order system

$$z'+Pz=0, \qquad P=(p_{ij})_{i,j=1}^{n}. \tag{2}$$

A nontrivial solution vector $z=\mathrm{col}(z_1,\cdots,z_n)$ of (2) is said to be *oscillatory* on an interval I if $z_k(x_k)=0$ for some $x_k\in I$, $k=1,\cdots,n$. The system (2) is said to be *oscillatory* if it has at least one oscillatory solution vector; otherwise, it is said to be *nonoscillatory* on I. The first-order system (2) is said to be *suborthogonal* on I if, for any nontrivial solution vector z and for any pair of points $s,t\in I$, the inner product $(z(s),z(t))\equiv\sum_{k=1}^{n}z_k(s)z_k(t)>0$ [13].

Suppose that $y=\mathrm{col}(y_1,\cdots,y_n)$ is a solution of (1). If we put $w=\mathrm{col}(y_1,\cdots,y_n,y_1',\cdots,y_n')$, the second-order system reduces to the first-order system

$$w'+Cw=0, \qquad C=\begin{pmatrix}0 & -\mathrm{I}\\ A & 0\end{pmatrix}, \tag{3}$$

where I is the $n\times n$ identity matrix. If (1) has a nontrivial solution $y=\mathrm{col}(y_1,\cdots,y_n)$ such that $y(a)=y(b)=0$, $a,b\in I$, $a<b$, then $y_i(a)=y_i(b)=0$, $i=1,\cdots,n$, and by Rolle's theorem, $y_i'(x_i)=0$ for some x_i, $a<x_i<b$, $i=1,\cdots,n$. This means that every component of w has a zero on I, i.e., (3) is oscillatory on I. Hence, any nonoscillation theorems for (3) may be useful in establishing explicit disconjugacy criteria for (1).

*Received by the editors January 10, 1984, and in revised form November 12, 1984.

†Department of Applied Mathematics and Statistics, State University of New York, Stony Brook, New York 11794.

Indeed, a nonoscillation theorem for (3) leads to a condition guaranteeing that at least one of the functions $y_1,\cdots,y_n,\ y_1',\cdots,y_n'$ does not vanish on I if $y=\operatorname{col}(y_1,\cdots,y_n)$ is a nontrivial solution of (1). In particular, any nonoscillation theorems for (3) lead to right- and left-disfocality criteria of (1). Furthermore, right- and left-disfocality of (1) are also implied by the suborthogonality of (3).

In 1930 Morse [12] studied the selfadjoint second-order systems in the setting of the calculus of variations and obtained extensions of the classical Sturm separation and comparison theorems [16]. Following his pioneering work, many results of similar type appeared in the literature [15]. Recently, in a series of papers [1]–[4], Ahmad and Lazer investigated the system (1) concerning the sign properties of the components of extremal solutions and comparison theorems for disconjugacy. Their results involved conditions imposed on the elements a_{ij} of the coefficient matrix A, while the earlier results had been stated in terms commonly encountered in the variational problems such as the positive semidefiniteness of certain matrices. Keener and Travis [9] also studied the Sturmian properties of (1) using the theory of μ_0-positive operators defined on a Banach space.

The purpose of this paper is to establish disconjugacy criteria and comparison theorems involving disfocality for the second-order system (1). In §2 we shall prove two sufficient conditions for disconjugacy (Theorems 1 and 2), one of which was obtained by noting the connection between the systems (1) and (3). In §3 we discuss comparison and separation theorems for the right- and left-disfocality of (1) (Theorems 4 and 6). These comparison theorems are proved by generalizing the method of the Riccati equation first used by Hille for the second-order equations [8], which was later adapted to higher-order equations by Nehari [14] and others [7], [11].

2. Disconjugacy criteria. Let $y=\operatorname{col}(y_1,\cdots,y_n)$ be a solution of (1) and put

$$w=\operatorname{col}(w_1,\cdots,w_{2n})=\operatorname{col}(\varepsilon y_1,\cdots,\varepsilon y_n,\ y_1',\cdots,y_n'),$$

for some $\varepsilon>0$. Then

$$w'=Bw, \tag{4}$$

where

$$B=B(x,\varepsilon)=\begin{pmatrix} 0 & \varepsilon \mathrm{I} \\ -\varepsilon^{-1}A(x) & 0 \end{pmatrix},$$

and I is the $n\times n$ identity matrix. According to a result of Nehari [13], (4) is nonoscillatory on the interval $I=[a,b]$ if

$$\int_a^b \|B\|\,dx<\pi/2,$$

where $\|B\|$ denotes the matrix norm $\sup_{\|z\|=1}\|Bz\|$ and $\|z\|$ is the Euclidean norm of vector z. As was pointed out in the preceding section, the nonoscillation of (4) implies that every nontrivial solution y of (1) possesses the property that at least one of the $2n$ functions $y_1,\cdots,y_n,\ y_1',\cdots,y_n'$ does not vanish on I, i.e., (1) is a fortiori disconjugate and right- and left-disfocal on I.

Put $u=\operatorname{col}(w_1,\cdots,w_n)$ and $v=\operatorname{col}(w_{n+1},\cdots,w_{2n})$. Then

$$\|Bw\|^2=(Bw,Bw)=\varepsilon^2\|v\|^2+\varepsilon^{-2}\|Au\|^2$$

$$\leqq\varepsilon^2\|v\|^2+\varepsilon^{-2}\|A\|^2\|u\|^2$$

$$\leqq\max\left(\varepsilon^2,\varepsilon^{-2}\|A\|^2\right)\|w\|^2;$$

consequently,

$$\|B\|\leqq\max\left(\varepsilon,\varepsilon^{-1}\|A\|\right),$$

and we have proved the following result.

THEOREM 1. *Let a and b, $a<b$, be points of the interval I. If*

$$\int_a^b\max\left(\varepsilon,\varepsilon^{-1}\|A(x)\|\right)dx<\pi/2,$$

for some $\varepsilon>0$, then every nontrivial solution y of the second-order system (1) *has the property that at least one component of y or y' does not vanish on $[a,b]$; in particular,* (1) *is disconjugate and right- and left-disfocal on $[a,b]$.*

If $\|A(x)\|$ is constant, the inequality condition in the above theorem is equivalent to $\|A(x)\|<\pi^2/4(b-a)^2$.

Next we shall establish a disconjugacy criterion of a different nature. Here we shall be concerned with a symmetric interval of the type $(-c,c)$. This is not an essential restriction because disconjugacy and zero properties are preserved under a translation.

THEOREM 2. *The system* (1) *is disconjugate on $I=(-c,c)$ if*

$$\frac{1}{2}\left(c^2-|x|^2\right)\sum_{j=1}^{n}|a_{kj}(x)|\leqq 1,\qquad x\in(-c,c), \tag{5}$$

$k=1,\cdots,n$.

Proof. Suppose that (1) is not disconjugate on $(-c,c)$. Then there exist a nontrivial solution $y=\operatorname{col}(y_1,\cdots,y_n)$ and a pair of points a and b, $-c<a<b<c$, such that $y_k(a)=y_k(b)=0$, $k=1,\cdots,n$. For each component y_k which is twice continuously differentiable and vanishes at a and b on $(-c,c)$, it is easily confirmed that

$$y_k(x)=(b-x)\int_a^x(b-t)^{-2}\int_b^t(b-s)\,y_k''(s)\,ds\,dt$$

[5], [10]. Let

$$|y_k''(x_k)|=\max_{a\leqq x\leqq b}|y_k''(x)|,$$

$k=1,\cdots,n$. Then we have for $a\leqq x\leqq b$,

$$|y_k(x)|\leqq|y_k''(x_k)||b-x|\int_a^x|b-t|^{-2}\int_b^t|b-s||ds||dt|$$

$$=\frac{1}{2}|y_k''(x_k)||b-x||a-x|.$$

Since $|b-x||a-x|<(c^2-|x|^2)$, $a\leqq x\leqq b$, we get the inequality

$$|y_k(x)|\leqq\frac{1}{2}|y_k''(x_k)|\left(c^2-|x|^2\right),\qquad a\leqq x\leqq b, \tag{6}$$

where the strict inequality holds unless $|y_k''(x_k)|=0$, $k=1,\cdots,n$. Note that $|y_k''(x_k)|=0$ if and only if $y_k\equiv 0$, due to the condition $y_k(a)=y_k(b)=0$. If we put

$$|y_m''(x_m)|\equiv\max_{1\leqq j\leqq n}|y_j''(x_j)|,$$

then $|y_m''(x_m)|>0$ since $y\not\equiv 0$. From the system (1),

$$|y_m''(x)|\leqq\sum_{j=1}^{n}|a_{mj}(x)||y_j(x)|;$$

in particular, for $x=x_m$,

$$0<|y_m''(x_m)|\leqq\sum_{j=1}^{n}|a_{mj}(x_m)||y_j(x_m)|. \tag{7}$$

Hence, there must exist an l, $1\leqq l\leqq n$, for which $|a_{ml}(x_m)||y_l(x_m)|>0$. For such an l, $y_l(x)\not\equiv 0$ and $y_l''(x_l)\neq 0$; therefore,

$$|y_l(x)|<\frac{1}{2}|y_l''(x_l)|\left(c^2-|x|^2\right),\qquad a\leqq x\leqq b, \tag{6'}$$

from (6). Substituting (6) and (6′) with $x=x_m$ in (7), we get

$$\begin{aligned}|y_m''(x_m)|&<\frac{1}{2}\left(c^2-|x_m|^2\right)\sum_{j=1}^{n}|a_{mj}(x_m)||y_j''(x_m)|\\&\leqq\frac{1}{2}\left(c^2-|x_m|^2\right)|y_m''(x_m)|\sum_{j=1}^{n}|a_{mj}(x_m)|,\end{aligned}$$

i.e.,

$$1<\frac{1}{2}\left(c^2-|x_m|^2\right)\sum_{j=1}^{n}|a_{mj}(x_m)|,$$

contrary to (5). This completes the proof.

3. Comparison theorems. In this section we shall frequently consider an interval of the form $I=[a,\omega)$, where ω may or may not be finite. If the system (1) is not right [left]-disfocal on $[a,\omega)$, we define $\eta(A,b)[\phi(A,b)]$, $a\leqq b<\omega$, to be the infimum of c, $b\leqq c<\omega$, such that there exists a nontrivial solution y of (1) satisfying $y(b)=y'(c)=0$ [$y'(b)=y(c)=0$]. Then there exists a nontrivial solution y of (1) satisfying $y(b)=y'(\eta(A,b))=0$ [$y'(b)=y(\phi(A,b))=0$]. If, on the other hand, (1) is right [left]-disfocal on $[a,\omega)$, we put $\eta(A,b)=\omega[\phi(A,b)=\omega]$ for all b.

Let $y_j=\mathrm{col}(y_{1j},\cdots,y_{nj})$, $j=1,\cdots,n$, be the solution vectors of (1) satisfying

$$y_{ij}(b)=0,\qquad y_{ij}'(b)=\delta_{ij}, \tag{8}$$

$i,j=1,\cdots,n$, for a fixed b, $a\leqq b<\omega$. Put $Y=(y_{ij})_{i,j=1}^{n}$ and $S=Y(Y'^{-1})$. If the system (1) is right-disfocal on $[b,\omega)$, the determinant of Y' does not vanish on $[b,\omega)$; therefore, S is continuously differentiable on $[b,\omega)$, and

$$S'=Y'Y'^{-1}+Y(Y'^{-1})'=\mathrm{I}+Y(-Y'^{-1}Y''Y'^{-1})$$

$$=\mathrm{I}+Y\left[-Y'^{-1}(-AY)Y'^{-1}\right]=\mathrm{I}+SAS,$$

i.e., S satisfies the matrix-matrix equation

$$S'=\mathrm{I}+SAS, \qquad S(b)=0. \tag{9}$$

Conversely, we assert that $\det Y'\neq 0$ as long as (9) has a continuous solution. Since $\det Y'(x)$ is a continuous function of x and $\det Y'(b)=1$, $\det Y'$ does not vanish on some right neighborhood N of the point b; that is, Y' is invertible on N. Therefore, YY'^{-1} is defined on N; moreover, $S=YY'^{-1}$ due to the uniqueness of solutions of the initial value problem (9) [6]. Suppose that $\det Y'$ vanishes at some point: Let x_0, $b<x_0<\omega$, be the first point to the right of b at which $\det Y'$ vanishes, while S remains continuous at x_0. Then the span G of the row vectors of $Y'(x_0)$ is at most $n-1$ dimensional, and there exists a nontrivial vector $\alpha=(\alpha_1,\cdots,\alpha_n)$ which is orthogonal to G, i.e.,

$$\sum_{j=1}^{n}\alpha_j y'_{ij}(x_0)=0, \tag{10}$$

$i=1,\cdots,n$. On the interval $[b,x_0)$ we have $S=YY'^{-1}$, which may be written as $SY'=Y$; indeed, this equality holds on $[b,x_0]$, because S, Y and Y' are continuous on $[b,x_0]$. Hence, for a fixed $x\in[b,x_0]$, the row vectors $(y_{k1},\cdots,y_{kn})$, $k=1,\cdots,n$, of Y are contained in the span of the row vectors of Y'. In particular, for $x=x_0$, every row vector $(y_{k1}(x_0),\cdots,y_{kn}(x_0))$, $k=1,\cdots,n$, of $Y(x_0)$ is contained in G. Since the vector α is orthogonal to G, it is orthogonal to the row vectors of $Y(x_0)$:

$$\sum_{j=1}^{n}\alpha_j y_{ij}(x_0)=0, \qquad i=1,\cdots,n. \tag{11}$$

Putting

$$y=\sum_{j=1}^{n}\alpha_j y_j, \qquad y_j=\mathrm{col}(y_{1j},\cdots,y_{nj}),$$

$j=1,\cdots,n$, we see that y is a nontrivial solution of the system (1). However, (10) and (11) require that $y(x_0)=y'(x_0)=0$, which cannot be satisfied unless y is the trivial solution. This contradiction proves that $\det Y'$ cannot vanish at x_0.

Thus we have proved the following theorem.

THEOREM 3. *Let b be a point on the interval $I=[a,\omega)$. Every nontrivial solution y of* (1) *with $y(b)=0$ has the property that $y'\neq 0$ on $[b,c)$, $a\leqq b<c\leqq\omega$, if and only if the Riccati system* (9) *has a continuous solution on $[b,c)$.*

In order to obtain the desired comparison theorems, we shall study the differential system satisfied by the components s_{ij} of the matrix $S=(s_{ij})_{i,j=1}^{n}$ in (9). As a first step, we need to determine the growth properties of s_{ij} in the right neighborhood of the

point b. Since

$$\det Y' = \sum_{\pi \in P} (\operatorname{sgn} \pi)\, y'_{1\pi(1)} y'_{2\pi(2)} \cdots y'_{n\pi(n)},$$

where P is the set of all permutations of the integers between 1 and n, the cofactor of y'_{ij} may be obtained from the above formula by setting $y'_{ij}=1$ and $y'_{ik}=0$, $k=1,\cdots,$ $j-1$, $j+1,\cdots,n$. Evidently, the cofactor of y'_{ij} is the sum of the products of the form

$$\pm y'_{1k_1} \cdots y'_{i-1,k_{i-1}} y'_{i+1,k_{i+1}} \cdots y'_{nk_n}, \tag{12}$$

where $1 \leqq k_m \leqq n$, $k_m \neq j$, $m=1,\cdots,i-1$, $i+1,\cdots,n$. Hence, every term in the sum contains a factor y'_{rs}, $r \neq s$, unless $i=j$. If $i=j$, the sum contains exactly one term of the form (12) for which $k_m=m$, $m=1,\cdots,i-1$, $i+1,\cdots,n$, and all the other terms contain a factor y_{rs}, $r \neq s$. Therefore, the diagonal and the off-diagonal elements of the inverse Y'^{-1} are, respectively, unity and zero at b. Furthermore, we see from (8) that the diagonal elements of Y have zeros of order exactly 1 at b, while the off-diagonal elements have zeros of order at least 2 at b. Consequently, simple computations show that the diagonal elements of $S=YY'^{-1}$ have zeros of order exactly 1 at b and the off-diagonal elements have zeros of order at least 2 at b.

From (9) we have $s_{ii}(b)=0$, $s'_{ii}(b)=1$, which implies that $s_{ii}>0$ on $(b,b+\varepsilon)$ for some $\varepsilon>0$, $i=1,\cdots,n$. On the other hand, for the off-diagonal element s_{ij}, $i \neq j$,

$$s'_{ij}(x) = \sum_{k=1}^{n} \sum_{m=1}^{n} a_{km}(x) s_{ik}(x) s_{mj}(x),$$

where the term $a_{ij}(x)s_{ii}(x)s_{jj}(x)$ is easily seen to be dominating as $x \to b+$, provided $a_{ij}(b) \neq 0$. This is because the term $a_{ij}s_{ii}s_{jj}$ has a zero of order exactly 2 at b if $a_{ij}(b) \neq 0$, while the other terms have zeros of order at least 3 at b. In particular, if $a_{ij}(b)>0$, $i \neq j$, then $s'_{ij}(x)>0$ on $(b,b+\varepsilon)$ for some $\varepsilon>0$. Thus, we have the following lemma.

LEMMA 1. *Let b be a point on the interval $I=[a,\omega)$. Then the solution matrix S of* (9) *is positive on $(b,b+\varepsilon)$ for some $\varepsilon>0$ if $a_{ij}(b)>0$, $i \neq j$.*

The condition $a_{ij}(b)>0$, $i \neq j$, in the above lemma may be relaxed. In view of the zero properties of the components s_{ij} at b, we may merely require that a_{ij} be nonnegative in a right neighborhood of b and $a_{ij}(x)=O((x-b)^{\lambda})$ for some λ, $0 \leqq \lambda < 1$, as $x \to b+$, $i \neq j$.

Another result needed for our proof is an extended version of [14, Lemma 3.2], which follows when a few obvious changes are made in the original proof.

LEMMA 2 [14, Lemma 3.2]. *Let $P_r(w_1,\cdots,w_m,t)$ and $P_r^*(w_1,\cdots,w_m,t)$ be two sets of polynomials in the variables $w_1,\cdots,w_m$ whose coefficients are nonnegative and continuous on $[a,\omega)$, $r=1,\cdots,m$. Suppose that*

$$\int_a^t P_r(w_1,\cdots,w_m,s)\,ds \leqq \int_a^t P_r^*(w_1,\cdots,w_m,s)\,ds, \qquad t \in [a,\omega),$$

for any set of nonnegative continuous functions $w_1,\cdots,w_m$ defined on $[a,\omega)$, $r=1,\cdots,m$. If there exist nonnegative differentiable functions $W_1,\cdots,W_m$ defined on $[a,\omega)$ satisfying the inequality

$$W_r' \geqq P_r^*(W_1,\cdots,W_m,t), \qquad W_r(a)=\alpha_r \geqq 0,$$

$r=1,\cdots,m$, *then the differential system*

$$w_r'=P_r(w_1,\cdots,w_m,t), \qquad w_r(a)=\beta_r, \qquad \alpha_r\geqq\beta_r\geqq 0,$$

$r=1,\cdots,m$, *has a continuous solution* $(w_1,\cdots,w_m)$ *on* $[a,\omega)$ *and* $w_r(t)\leqq W_r(t)$, $r=1,\cdots,m$, *on* $[a,\omega)$.

Let $B=B(x)$ be an $n\times n$ matrix with real elements which are continuous on the x-interval $[a,\omega)$. We shall prove a comparison theorem for (1) and the system

$$y''+By=0, \qquad B=(b_{ij})_{i,j=1}^{n}, \tag{13}$$

by applying Lemma 2 to the corresponding Riccati systems

$$S'=\mathrm{I}+SAS, \qquad S(c)=\alpha\geqq 0, \qquad c\in[a,\omega), \tag{14}$$

$$T'=\mathrm{I}+TBT, \qquad T(c)=0, \qquad c\in[a,\omega). \tag{15}$$

THEOREM 4. *Let* $A=(a_{ij})_{i,j=1}^{n}$ *and* $B=(b_{ij})_{i,j=1}^{n}$ *be matrices with real elements which are continuous on an interval* $[a,\omega)$. *Assume that* $0\leqq b_{ij}\leqq a_{ij}$, $i,j=1,\cdots,n$, *on* $[b,\omega)$ *for some* b, $a\leqq b<\omega$, *and that* $a_{ij}(b)>0$, $i\neq j$. *Then* $\eta(B,c)\geqq\eta(A,b)$, $a\leqq b\leqq c<\omega$.

Proof. If y is a nontrivial solution of the system (1) satisfying $y(b)=0$, then $y'(x)\neq 0$, $b\leqq x<\eta(A,b)$. Thus, the Riccati system (9) has a continuous solution S on $[b,\eta(A,b))$ by Theorem 3. The solution S is positive on $(b,b+\varepsilon)$ for some $\varepsilon>0$ by Lemma 1; therefore, S is positive throughout the interval $(b,\eta(A,b))$ due to the nonnegativity of the coefficients of the system (9). Hence, on the interval $[c,\eta(A,b))$, $b\leqq c<\eta(A,b)$, S is continuous and satisfies (14) with $\alpha=S(c)$. From Lemma 2 and the inequalities $0\leqq b_{ij}\leqq a_{ij}$, $i,j=1,\cdots,n$, we conclude that (15) possesses a nontrivial solution T which is continuous on $[c,\eta(A,b))$. It follows finally from Theorem 3 that the system (13) is right-disfocal on $[c,\eta(A,b))$, i.e., $\eta(B,c)\geqq\eta(A,b)$.

If $\eta(A,b)\leqq c<\omega$, the inequality holds trivially.

We remark that the following "separation theorem" results when $A=B$ in Theorem 4: Let $A=(a_{ij})_{i,j=1}^{n}$ be a matrix with real elements which are continuous on $[a,\omega)$. Assume that $a_{ij}\geqq 0$, $i,j=1,\cdots,n$, on some interval $[b,\omega)$, $a\leqq b<\omega$, and that $a_{ij}(b)>0$, $i\neq j$. If $b\leqq x_1<x_2<\eta(A,b)$, then the system (1) has no nontrivial solution y such that $y(x_1)=y'(x_2)=0$, i.e., $\eta(A,c)\geqq\eta(A,b)$, $b\leqq c<\omega$.

There are parallel results for left-disfocality, which we summarize below. Let $u_j=\mathrm{col}(u_{1j},\cdots,u_{nj})$, $j=1,\cdots,n$, be the solutions of (1) satisfying

$$u_{ij}(b)=\delta_{ij}, \qquad u_{ij}'(b)=0,$$

for some fixed point b, $a\leqq b<\omega$, $i,j=1,\cdots,n$. Put $U=(u_{ij})_{i,j=1}^{n}$ and $V=-U'U^{-1}$. If the system (1) is left-disfocal on $[b,\omega)$, $\det U$ does not vanish on $[b,\omega)$. Hence, V is continuously differentiable on $[b,\omega)$ and

$$V'=A+V^2, \qquad V(b)=0. \tag{16}$$

Conversely, $\det U$ does not vanish on $[b,\omega)$ if (16) has a continuous solution V on $[b,\omega)$. The proof is omitted because it is similar to the preceding case of right-disfocality. To recapitulate, we have

THEOREM 5. *Let* b *be a point on the interval* $I=[a,\omega)$. *Every nontrivial solution* y *of* (1) *with* $y'(b)=0$ *does not vanish on* $[b,c)$, $a\leqq b<c\leqq\omega$, *if and only if the Riccati system* (16) *has a continuous solution on* $[b,c)$.

Due to the form of the Riccati system (16)—the coefficient matrix A appears by itself—the proof of the comparison theorem for this case is substantially simpler, and the resulting theorem involves inequality conditions imposed on the integrals of the coefficients, rather than on the coefficients themselves.

THEOREM 6. *Let* $A=(a_{ij})_{i,j=1}^{n}$ *and* $B=(b_{ij})_{i,j=1}^{n}$ *be matrices with nonnegative, real elements which are continuous on* $[a,\omega)$. *If* $a_{ij}(b)>0$ *and*

$$\int_c^t b_{ij}(s)\,ds \leqq \int_c^t a_{ij}(s)\,ds, \qquad c \leqq t < \omega, \tag{17}$$

$i,j=1,\cdots,n$, *for some* b *and* c, $a \leqq b \leqq c < \omega$, *then* $\phi(B,c) \geqq \phi(A,b)$.

Proof. By Theorem 5, the Riccati system has a continuous solution V on $[b,\phi(A,b))$. Since $V(b)=0$ and $V'(b)=A(b)>0$, $V>0$ on $(b,b+\varepsilon)$, for some $\varepsilon>0$, and therefore $V>0$ on $(b,\phi(A,b))$ due to the nonnegativity of a_{ij}. If $\phi(A,b) \leqq c$, then the inequality $\phi(B,c) \geqq \phi(A,b)$ holds trivially. If $b \leqq c < \phi(A,b)$, the solution V satisfies

$$V' = A + V^2, \qquad V(c) \geqq 0.$$

From (17) and Lemma 2, we deduce that the system

$$W' = B + W^2, \qquad W(c) = 0,$$

has a continuous solution W on $[c,\phi(A,b))$. Therefore, according to Theorem 5, every nontrivial solution y of the second-order system $y''+By=0$, $y'(c)=0$, is such that $y \neq 0$ on $[c,\phi(A,b))$, i.e., $\phi(B,c) \geqq \phi(A,b)$.

By putting $A=B$ in the above comparison theorem, we again obtain a separation theorem. However, all we can conclude in this case is that the system (1) has no nontrivial solution y such that $y'(c)=y(d)=0$, $b \leqq c < d < \phi(A,b)$, where c is the lower limits of the integrals in (17). If (17) holds for all c, $b \leqq c < \omega$—as in the case $b_{ij} \leqq a_{ij}$, $i,j=1,\cdots,n$—then the above separation theorem is valid for any two points b and c, $b \leqq c < d < \phi(A,b)$.

REFERENCES

[1] S. AHMAD AND A. C. LAZER, *On the components of extremal solutions of second order systems*, this Journal, 8 (1977), pp. 16–23.

[2] ———, *A new generalization of the Sturm comparison theorem to selfadjoint systems*, Proc. Amer. Math. Soc., 68 (1978), pp. 185–188.

[3] ———, *An N-dimensional extension of the Sturm separation and comparison theory to a class of non-selfadjoint systems*, this Journal, 9 (1978), pp. 1137–1150.

[4] ———, *On an extension of Sturm's comparison theorem to a class of nonselfadjoint second-order systems*, Nonlinear Anal. Theory, Meth., Appl., 4 (1980), pp. 497–501.

[5] G. A. BESSMERTNYH AND A. YU. LEVIN, *Some inequalities satisfied by differentiable functions of one variable*, Dokl. Akad. Nauk SSSR, 144 (1962), pp. 471–474; Sov. Math. Dokl., 3 (1962), pp. 737–740.

[6] G. BIRKHOFF AND G. C. ROTA, *Ordinary Differential Equations*, 3rd ed., John Wiley, New York, 1978.

[7] G. J. BUTLER AND L. H. ERBE, *Nonlinear integral Riccati systems and comparison theorems for linear differential equations*, this Journal, 14 (1983), pp. 463–473.

[8] E. HILLE, *Non-oscillation theorems*, Trans. Amer. Math. Soc., 64 (1948), pp. 234–252.

[9] M. S. KEENER AND C. C. TRAVIS, *Sturmian theory for a class of nonselfadjoint differential systems*, Ann. Mat. Pura Appl. (Series 4), 123 (1980), pp. 247–266.

[10] W. J. KIM, *On a theorem of Pokornyi*, Proc. Amer. Math. Soc., 23 (1969), pp. 343–346.
[11] _____, *Generalized comparison theorems for disfocality types of the equation* $L_n y + py = 0$, J. Math. Anal. Appl., 109 (1985), pp. 182–193.
[12] M. MORSE, *A generalization of the Sturm separation and comparison theorems in n-space*, Math. Ann., 103 (1930), pp. 52–69.
[13] Z. NEHARI, *Oscillation theorems for systems of linear differential equations*, Trans. Amer. Math. Soc., 139 (1969), pp. 339–347.
[14] _____, *Nonlinear techniques for linear oscillation problems*, Trans. Amer. Math. Soc., 210 (1975), pp. 387–406.
[15] W. T. REID, *Ordinary Differential Equations*, John Wiley, New York, 1971.
[16] J. C. F. STURM, *Mémoire sur les équations différentielles linéaires de second ordre*, J. Math. Pures Appl., 1 (1836), pp. 106–186.

SIAM J. MATH. ANAL.
Vol. 17, No. 5, September 1986

006

EVOLUTION PROBLEMS WITH HYSTERESIS IN THE SOURCE TERM*

A. VISINTIN†

Abstract. Jump conditions having a hysteresis effect are introduced to model some inertial mechanisms which yield criteria of control of systems. Examples can be found in biology, chemistry, engineering, and economics.

Both orientations of the hysteresis loops are considered, corresponding to "delay" and "anticipation" between input and output variables.

Existence results are proved for parabolic and hyperbolic problems and also for the Stefan problem with a term of this type in the second member. A regularizing effect is exhibited by the "jump with delay" condition.

Key words. variational problem, hysteresis, existence results

1. Introduction.

1) We introduce a simple *hysteresis* relationship between two functions of time: $u(t)$, $w(t)$. Fix ρ_1, $\rho_2 \in \mathbb{R}$ with $\rho_1 < \rho_2$. Let $u \in C^0([0,T])$ and w^0 be given such that

$$\begin{array}{ll} \text{if } u(0) \leqq \rho_1, & \text{then } w^0 = -1; \\ \text{if } u(0) \geqq \rho_2, & \text{then } w^0 = 1; \\ \text{if } \rho_1 < u(0) < \rho_2, & \text{then } w^0 = -1 \text{ or } 1. \end{array} \tag{1.1}$$

We say that the function w: $[0,T] \to \{-1,1\}$ fulfills a *jump condition with delay* if

$$\begin{array}{l} w(0) = w^0; \\ \text{if } u(t) \leqq \rho_1 \ (u(t) \geqq \rho_2, \text{ respectively}) \text{ then } w(t) = -1 \ (w(t) = 1, \text{ respectively}) \\ \forall t \in [0,T]; \\ w \text{ can jump from } -1 \text{ to } 1 \text{ (from } 1 \text{ to } -1, \text{ respectively) at time } t \text{ only if} \\ u(t) = \rho_2 \ (\rho_1, \text{ respectively}), \text{ these are the only discontinuities of } w. \end{array} \tag{1.2}$$

This means that for any $t \in [0,T[$ if $w(t) = -1$ ($w(t) = 1$, respectively) then w remains constant for $\tau > t$ as long as $u(\tau) < \rho_2$ ($u(\tau) > \rho_1$, respectively); if u reaches ρ_2 (ρ_1, respectively), then w jumps to 1 (-1, respectively), where it remains till u reaches ρ_1 (ρ_2, respectively), and so on. Here delay is meant with respect to the usual condition taking place in any fixed point of $[\rho_1, \rho_2]$: the jump is delayed w.r.t. u, not w.r.t. time.

2) Examples of hysteresis relations of this type are quite common in technology, for instance a thermostat which is switched on or off according to temperature, but with an inertial behavior in its dependence on the latter, so that actually the switching off value ρ_2 is greater than the swtiching on one ρ_1. A different model for problems with thermostats has been studied by Glashoff and Sprekels in [1], [2].

Another example is given by an irrigation model; here u represents water saturation, $w+1$ is the intensity of a water source due to irrigation and (1.2) corresponds to a

* Received by the editors December 7, 1982, and in final revised form November 15, 1984. This work has been supported by the Deutsche Forschungsgemeinschaft through the Sonderforschungsbereich 123.

† Institut für Angewandte Mathematik, Universität Heidelberg, Im Neuenheimer Feld 294, 6900 Heidelberg 1, West Germany. Permanent address: Istituto di Analisi Numerica del C.N.R., c.so C. Alberto 5, 27100-Pavia, Italy.

control criterion for switching the water source either on or off according to the water saturation (see [9, p. 488]).

Further examples arise in biology and chemistry, as in the model studied by Hoppensteadt and Jäger in [3], [4] and treated here in §6.

Mathematical models for hysteresis phenomena have been studied by Krasnosel'skiĭ and Pokrovskiĭ (see [5], [6], e.g.) and by the present author (see [10]–[13]).

3) It can be useful to compare (1.2) with the usual jump relationship. The latter is monotone, but its graph is not maximal monotone; thus the graph of the corresponding operator $u \mapsto w = \operatorname{sign}(u)$ is not closed with respect to the strong topology of $C^0([0,T])$ for u and the weak star topology of $L^\infty(0,T)$ for w. Consequently one is induced to work with the closure of this graph, that is to replace the sign function by the sign graph.

Similarly the functional $u \mapsto w$ defined by (1.2) (for any compatible w^0) is not closed with respect to the strong topology of $C^0([0,T])$ for u and the weak star topology of $L^\infty(0,T)$ for w. Therefore we shall study the form of its closure with respect to these topologies.

The couples $(u,w) \in C^0([0,T]) \times L^\infty(0,T)$ which belong to this closure fulfill the following conditions:

(1.3)
$-1 \leqq w(t) \leqq 1$;
if $u(t) < \rho_1$ ($u(t) > \rho_2$, respectively), then $w(t) = -1$ ($w(t) = 1$, respectively);
if $\rho_1 < u(t) < \rho_2$, then w is constant in a neighborhood of t;
if $u(t) = \rho_1$ ($u(t) = \rho_2$, respectively), then $w(t)$ is nonincreasing (nondecreasing, respectively) in a neighborhood of t (see Fig. 1).

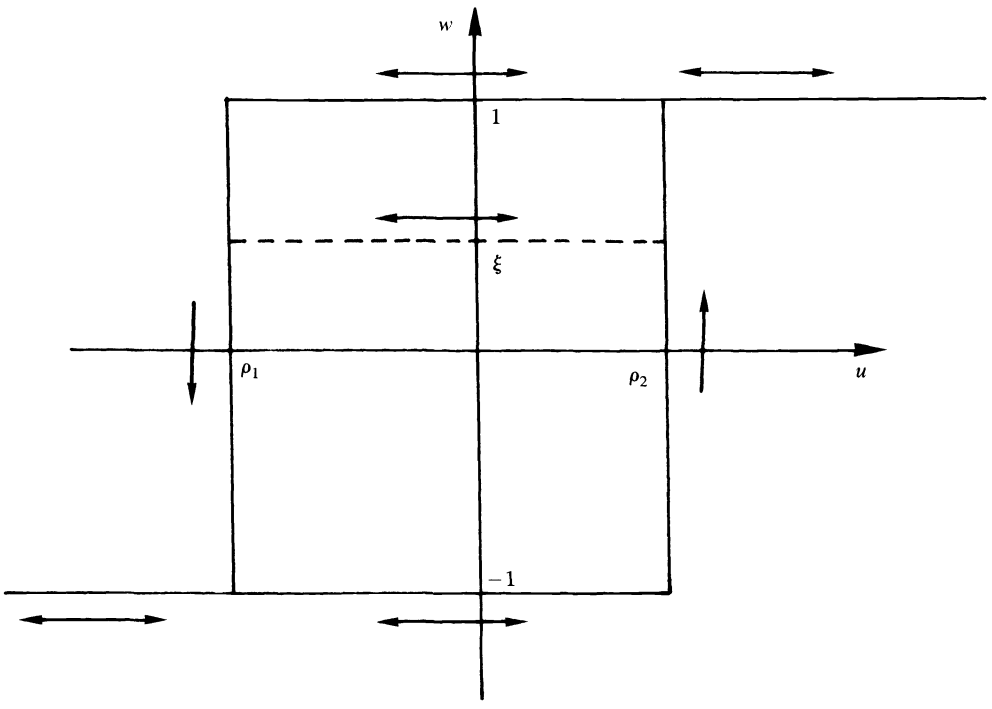


FIG. 1. *Arrows indicate the direction of movement of* $(u(t), w(t))$ *as* t *increase.* $\xi \in [-1,1]$ *is generic.*

Indeed by means of small perturbations of $u(t)$ where this reaches ρ_1 (ρ_2, respectively) as a local minimum (maximum, respectively), we get perturbed w's which attain the constant values -1 or 1 as long as $\rho_1 < u(t) < \rho_2$; then by weak star convergence in $L^\infty(0,T)$ we can get any w which attains arbitrary constant values betweeen -1 and 1 as long as $\rho_1 < u(t) < \rho_2$.

We note that any w fulfilling (1.2) has bounded total variation in $[0,T]$: $w \in BV(0,T)$; thus the initial condition is meaningful and $w' \in C^0([0,T])$. Indeed w has a variation equal to 2 just when u reaches one of the thresholds ρ_1, ρ_2 and moreover the uniformly continuous function u can have just a finite number of oscillations between ρ_1 and ρ_2. Also any w fulfilling (1.3) has bounded total variation in $[0,T]$.

We introduce some notation:

$$\alpha(\xi) \equiv (\xi-\rho_2)^+ - (\xi-\rho_1)^-, \quad \beta(\xi) \equiv \xi - \alpha(\xi) \quad \forall \xi \in \mathbb{R}; \tag{1.4}$$

$$S(\eta) \equiv \begin{cases} \{-1\} & \text{if } \eta < 0, \\ [-1,1] & \text{if } \eta = 0 \\ \{1\} & \text{if } \eta > 0, \end{cases} \quad \forall \eta \in \mathbb{R}; \tag{1.5}$$

$$R(\eta) \equiv \begin{cases}]-\infty, 0[& \text{if } \eta = \rho_1, \\ \{0\} & \text{if } \rho_1 < \eta < \rho_2 \\ [0, +\infty[& \text{if } \eta = \rho_2, \end{cases} \quad \forall \eta \in [\rho_1, \rho_2]. \tag{1.6}$$

It is easy to check that (1.3) entails

$$\int_0^T w[\alpha(u) - v]\, dt \geqq \int_0^T (|\alpha(u)| - |v|)\, dt \quad \forall v \in L^1(0,T), \tag{1.7}$$

i.e.

$$w \in S(\alpha(u)) \quad \text{in }]0,T[. \tag{1.8}$$

Conditions (1.3) imply also

$$_{(C^0([0,T]))'}\langle w', \beta(u) - v \rangle_{C^0([0,T])} \geqq 0 \quad \forall v \in C^0([0,T]) \text{ such that } \rho_1 \leqq v \leqq \rho_2, \tag{1.9}$$

i.e.,

$$w' \in R(\beta(u)) \quad \text{in } (C^0([0,T]))'; \tag{1.10}$$

the latter is equivalent to

$$w' \in S^{-1}\left(\frac{2}{\rho_2 - \rho_1} \left[\beta(u) - \frac{\rho_1 + \rho_2}{2} \right] \right), \quad \text{i.e.} \quad \frac{2}{\rho_2 - \rho_1} \left[\beta(u) - \frac{\rho_1 + \rho_2}{2} \right] \in S(w'), \tag{1.11}$$

which is to be understood as follows

$$\frac{2}{\rho_2 - \rho_1} {}_{C^0([0,T])}\left\langle \beta(u) - \frac{\rho_2 + \rho_1}{2}, w' - v \right\rangle_{(C^0([0,T]))'} \geqq \int_0^T (|w'| - |v|) \quad bav \in (C^0([0,T]))' \tag{1.12}$$

(here $\int_0^T |v|$ denotes the total mass of v in $[0,T]$); this can be easily deduced by (1.3). (We point out that calculations are simplified in the case of $\rho_1 = -1$, $\rho_2 = 1$.) (1.7) and (1.9) express the property (1.3) in variational form. We weaken the compatibility

condition (1.1) requiring just that

$$\text{(1.13)}\qquad \begin{aligned} &\text{if } u(0)<\rho_1 \ (u(0)>\rho_2, \text{ respectively}), \text{ then } w^0=-1 \ (w^0=1, \text{ respectively});\\ &\text{if } \rho_1\leqq u(0)\leqq\rho_2, \text{ then } -1\leqq w(0)\leqq 1; \end{aligned}$$

then for any compatible couple $(u,w^0)\in C^0([0,T])\times[-1,1]$ we consider the set of the $w\in BV(0,T;[-1,1])$ such that $w(t)\to w^0$ as $t\to 0^+$, (1.7) and (1.9) hold.

For any fixed w^0 this multivalued correspondence $u\mapsto w$ is closed with respect to the strong topology of $C^0([0,T])$ for u and the weak star topology of $BV(0,T)$ for w; hence it is the closure of the functional defined by (1.2).

So far we have assumed the couple (ρ_1,ρ_2) to be fixed; an interesting generalization is obtained as follows. Set $\mathscr{P}\equiv\{\rho=(\rho_1,\rho_2)\in\mathbb{R}^2\mid\rho_1<\rho_2\}$; let $u\in C^0([0,T])$ and for any ρ let w_ρ denote the function w defined by (1.2) and corresponding to an initial datum w_ρ^0 compatible with u in the sense of (1.1); let μ be a measure over $\mathscr{P}$; then we can consider the functional $u\mapsto w=\int_{\mathscr{P}} w_\rho\,d\mu_\rho$; this corresponds to the classical Preisach model introduced for ferromagnetism and has been studied in [12], for example,

4) After considering hysteresis (delay), we take into account the inverse effect, namely *anticipation*.

Let u: $[0,T]\to\mathbb{R}$ be absolutely continuous; we shall say that a measurable function $w(t)$ fulfills a *jump condition with anticipation* if a.e. in $]0,T[$

$$\text{(1.14)}\qquad \begin{aligned} &\text{if } u(t)<\rho_1 \ (u(t)>\rho_2, \text{ respectively}) \text{ then } w(t)=-1 \ (w(t)=1, \text{ respectively}),\\ &\text{if } \rho_1\leqq u(t)\leqq\rho_2, \text{ then } w(t)\in S(u'(t)) \text{ (see Fig. 2)} \end{aligned}$$

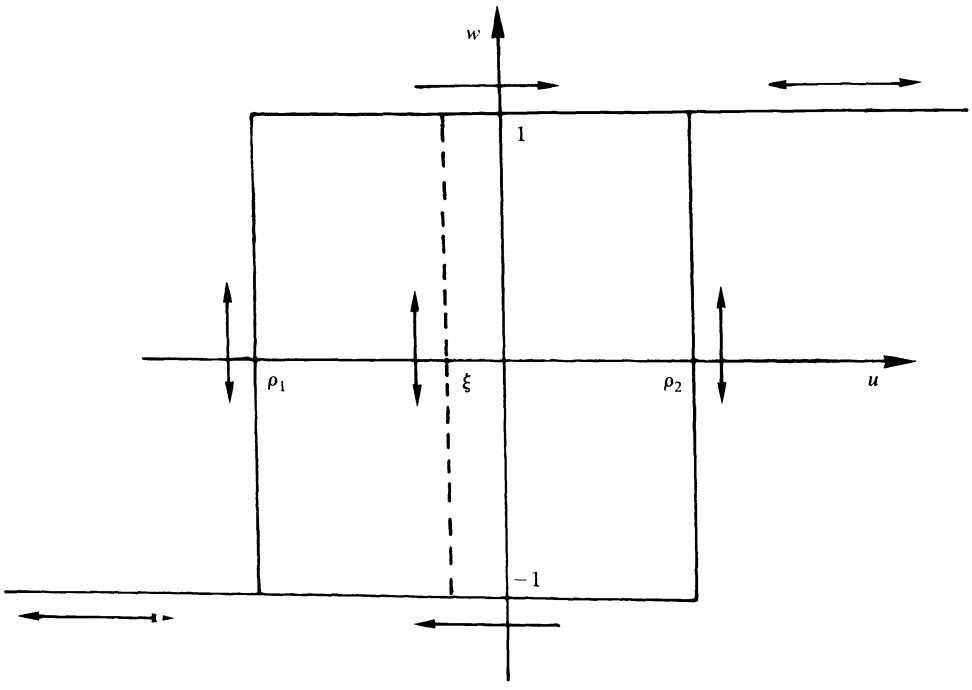


FIG. 2. *Arrows indicate the direction of movement of* $(u(t), w(t))$ *as* t *increases.* $\xi\in[\rho_1,\rho_2]$ *is generic.*

or equivalently

$$\begin{aligned} &w\in S(\alpha(u)) \quad \text{a.e. in }]0,T[,\\ &w\in S\big(\beta(u)'\big) \quad \text{a.e. in }]0,T[\end{aligned} \tag{1.15}$$

or also

$$w\in S\big(\alpha(u)+\beta(u)'\big) \quad \text{a.e. in }]0,T[. \tag{1.16}$$

Note that here no initial condition is required for w; moreover the graph of the multivalued functional $u\mapsto w$ is closed with respect to the strong topology of $W^{1,1}(0,T)$ for u and the weak star topology of $L^\infty(0,T)$ for w.

If $\rho_1=-1$ and $\rho_2=1$, then the behavior of $(u(t),w(t))$ in (1.7), (1.9) and in (1.14) are reciprocally inverse in $[-1,1]^2$; in particular the esterior loop is covered with opposite orientations in the two cases.

5) Anticipation corresponds to the case in which in a certain range the output w depends on the trend of the input u, not on its value. Examples of this situation can be found in sociology, economics, and so on. For instance delay and anticipation in the sense introduced above are possible attitudes of economical operators; here w may represent a decision ($w=1$ for Yes, $w=-1$ for No, say), whereas u may be a measure of the advantages offered by the operation under consideration.

In this last case and in the example of irrigation the dependence $u\mapsto w$ corresponds to a criterion of control, whereas in the example of thermostat it describes the behavior of a system. In some cases, for instance that of the economical operator, it may be possible to choose between the strategy of delay and that of anticipation (and also to select the thresholds ρ_1, ρ_2). If the input u is continuous in time, then the delay criterion offers the advantage of bounding the total variation of the output w; for instance in the model of irrigation the number of interventions on the irrigation system will be finite. On the other hand, if the input is absolutely continuous in time, then the anticipation criterion allows to take advantage of the forecast offered to some extent by the inertia of the variable u.

6) Let D be a bounded domain of $\mathbb{R}^n$ ($N\geqq 1$). We shall deal with PDE's of the form

$$u_t-\Delta u+w-f \quad \text{in } Q\equiv D\times]0,T[; \tag{1.17}$$

$$\begin{aligned} &u_t-\Delta u=g && \text{in } Q,\\ &\frac{\partial u}{\partial \nu}+w=h && \text{on } \Sigma\equiv \partial D\times]0,T[; \end{aligned} \tag{1.18}$$

$$u_{tt}-\Delta u+w=f \quad \text{in } Q; \tag{1.19}$$

where f, g, h are data, $\Delta\equiv\sum_{i=1}^N \partial^2/\partial x_i^2$ and $\partial/\partial\nu$ denotes the outward normal derivative.

Each of these equations will be coupled with either the hysteresis or the anticipation relationships introduced above. We also consider a Stefan problem with hysteresis in the source; this corresponds to coupling the hysteresis relation with the system

$$\begin{aligned} &(u+\lambda\chi)_t-\Delta u+w=f && \text{in } Q;\\ &\chi\in H(u) && \text{in } Q, \end{aligned} \tag{1.20}$$

where λ is a positive constant representing latent heat and H denotes the Heaviside graph ($H(\xi)=\{0\}$ if $\xi<0$, $H(0)=[0,1]$, $H(\xi)=\{1\}$ if $\xi>0$).

The present author has already studied hysteresis in connection with PDE's (see [10], [11], [12] and [13] for a survey). In particular in [11] existence and approximation results were proved for (1.7), (1.9) coupled with the equation

$$(u+w)_t-\Delta u=f \quad \text{in } Q, \tag{1.21}$$

with suitable initial and boundary condition. This setting generalizes the weak formulation of the Stefan problem.

Here we shall prove existence results for (the weak formulations of) equations (1.17), (1.18) and (1.19), each one coupled with the hysteresis relation (1.7), (1.9) (problems (P1), (P2), (P3), respectively; §2), for the same equations coupled with the anticipation relation (1.15) (problems (P4), (P5), (P6), respectively; §3) and for the Stefan problem with hysteresis in the source, i.e. (1.20) coupled with (1.7), (1.9) (problem (P7), §5). We also give several complementary results.

Finally in §6 we consider the problem studied by Hoppensteadt and Jäger in [3], [4]. There a system of two diffusion equations is coupled with a hysteresis relationship of the form (1.2). We replace the latter by the weaker condition (1.3), i.e. with the system (1.7), (1.9), and prove existence of a variational solution of the problem.

We stress that the formulation of the hysteresis relation we give here is different from that used by Jäger in [4].

For all of these problems the uniqueness of the solution is an open question. We are just able to prove the weaker result of Proposition 8.

2. Jump with delay.

1) Set $V=H_0^1(D)$, Hilbert space with norm $\|v\|_V=(\int_D|\overline{\nabla}v|^2dx)^{1/2}$, and $A\colon V\to V'$ defined by ${}_{V'}\langle Au,v\rangle_V=\int_D\overline{\nabla}u\cdot\overline{\nabla}v\,dx\ \forall u,\ v\in V$. Let

$$f\in L^2(0,T,V'); \tag{2.1}$$

$$u^0\in L^2(D) \text{ and } w^0\in L^\infty(D), \text{ such that } w^0\in S(\alpha(u^0)) \text{ a.e. in } D. \tag{2.2}$$

We introduce a weak formulation of (1.17), (1.3).

(P1) Find $u\in L^2(0,T,V)\cap H^1(0,T,V')$ ($\subset C^0([0,T];L^2(D))$) and $w\in L^\infty(Q)$ such that $\beta(u(x,\cdot))\in C^0([0,T])$, $w(x,\cdot)\in BV(0,T)$ a.e. in D and

$$w\in S(\alpha(u)) \quad \text{a.e. in } Q, \tag{2.3}$$

$$({}_{C^0([0,T])'}\langle w_t,\beta(u)-v\rangle_{C^0([0,T])}\geqq 0 \quad \forall v\in C^0([0,T]) \text{ such that } \rho_1\leqq v\leqq\rho_2, \text{ a.e. in } D, \tag{2.4}$$

$$\lim_{t\to 0^+}[w(x,t)-w^0(x)]\cdot[\beta(u^0(x))-v]\geqq 0 \quad \forall v\in[\rho_1,\rho_2], \text{ a.e. in } D, \tag{2.5}$$

$$u_t+Au+w=f \quad \text{in } V', \quad \text{a.e. in }]0,T[, \tag{2.6}$$

$$u(x,0)=u^0(x) \quad \text{a.e. in } D. \tag{2.7}$$

THEOREM 1. *Assume that* (2.2) *holds and that*

$$u^0 \in V; \tag{2.8}$$

$$f = f_1 + f_2, \quad f_1 \in L^2(Q), \quad f_2 \in W^{1,1}(0,T;V'). \tag{2.9}$$

Then problem (P1) *has at least one solution such that moreover*

$$u \in H^1(0,T;L^2(D)) \cap C_S^0([0,T];V); \tag{2.10}$$

$$w \in L^2(D; BV(0,T)). \tag{2.11}$$

We recall the definition of the space of scalarly continuous functions with values in a Banach space B:

$$C_S^0([0,T];B) = \{ v\colon [0,T] \to B \mid \forall \varphi \in B',\ t \mapsto {}_B\langle v(t), \varphi \rangle_{B'} \text{ is continuous in } [0,T] \}. \tag{2.12}$$

Throughout this paper u and w are regarded as functions of t with values into a space of functions of x, as well as functions of x with values into a space of functions of t.

Proof. (i) *Approximation.* Let $m \in \mathbb{N}$, $k = T/m$. Set

$$\begin{aligned} &f_m^n = f_{1m}^n + f_{2m}^n, \qquad f_{1m}^n(x) = \frac{1}{k} \int_{(n-1)k}^{nk} f_1(x,t)\,dt \quad \text{a.e. in } D; \\ &f_{2m}^n = f_2(nk) \quad \text{in } V', \text{ for } n = 1, \cdots, m; \end{aligned} \tag{2.13}$$

$$K(\xi,\eta) = \begin{cases} \{-1\} & \text{if } \xi < \rho_1 \\ [-1,\eta] & \text{if } \xi = \rho_1 \\ \{\eta\} & \text{if } \rho_1 < \xi < \rho_2 \\ [\eta,1] & \text{if } \xi = \rho_2 \\ \{1\} & \text{if } \xi > \rho_2 \end{cases} \quad \forall \xi \in \mathbb{R}, \forall \eta \in [-1,1]. \tag{2.14}$$

$(\mathrm{P1})_m$ Find $u_m^n \in V$, $w_m^n \in L^\infty(D)$ for $n = 1, \cdots, m$, such that, setting $u_m^0 = u^0$, $w_m^0 = w^0$ a.e. in D,

$$\frac{u_m^n - u_m^{n-1}}{k} + Au_m^n + w_m^n = f_m^n \quad \text{in } V', \text{ for } n = 1, \cdots, m, \tag{2.15}$$

$$w_m^n \in K(u_m^n, w_m^{n-1}) \quad \text{a.e. in } D, \text{ for } n = 1, \cdots, m. \tag{2.16}$$

$\forall m \in \mathbb{N}$ $(\mathrm{P1})_m$ can be solved step by step. Fix $n \in \{1, \cdots, m\}$ and assume that u_m^{n-1}, w_m^{n-1} are known. $K(\cdot, w_m^{n-1}(x))$ is a maximal monotone graph a.e. in D. Therefore (2.15) is equivalent to the minimization of a coercive, strictly convex, lower semicontinuous functional $V \to \mathbb{R}$; this problem has a unique solution u_m^n (which can be approximated by standard space-discretization methods).

(ii) *Estimates.* Take a generic $l \in \{1, \cdots, m\}$, multiply (2.15) by $u_m^n - u_m^{n-1}$ and sum for $n = 1, \cdots, l$. Notice that

$$\sum_{n=1}^{l} \int_D \frac{u_m^n - u_m^{n-1}}{k} \left(u_m^n - u_m^{n-1}\right) dx = k \sum_{n=1}^{l} \left\| \frac{u_m^n - u_m^{n-1}}{k} \right\|^2_{L^2(D)}, \tag{2.17}$$

$$\sum_{n=1}^{l} \int_D w_m^n \left(u_m^n - u_m^{n-1}\right) dx \geqq -k \sum_{n=1}^{l} \left\| \frac{u_m^n - u_m^{n-1}}{k} \right\|_{L^1(D)}, \tag{2.18}$$

$$\begin{aligned} \sum_{n=1}^{l} {}_{V'}\langle A u_m^n, u_m^n - u_m^{n-1} \rangle_V &= \sum_{n=1}^{l} \int_D \overline{\nabla} u_m^n \cdot \overline{\nabla} \left(u_m^n - u_m^{n-1}\right) dx \\ &\geqq \frac{1}{2} \sum_{n=1}^{l} \left(\|u_m^n\|_V^2 - \|u_m^{n-1}\|_V^2 \right) = \frac{1}{2} \|u_m^l\|_V^2 - \frac{1}{2} \|u^0\|_V^2, \end{aligned} \tag{2.19}$$

$$\sum_{n=1}^{l} \int_D f_{1m}^n \left(u_m^n - u_m^{n-1}\right) dx \leqq \|f_1\|_{L^2(Q)} \left(k \sum_{n=1}^{l} \left\| \frac{u_m^n - u_m^{n-1}}{k} \right\|^2_{L^2(D)} \right)^{1/2}, \tag{2.20}$$

$$\begin{aligned} &\sum_{n=1}^{l} {}_{V'}\langle f_{2m}^n, u_m^n - u_m^{n-1} \rangle_V \\ &\quad = {}_{V'}\langle f_{2m}^n, u_m^l \rangle_V - {}_{V'}\langle f_{2m}^1, u^0 \rangle_V - \sum_{n=1}^{l} {}_{V'}\langle f_{2m}^n - f_{2m}^{n-1}, u_m^{n-1} \rangle_V \\ &\quad \leqq \text{Const.} \|f_{2m}\|_{W^{1,1}(0,T;V')} \max_{n=0,\cdots,l} \|u_m^n\|_V. \end{aligned} \tag{2.21}$$

Using Gronwall's lemma we get

$$k \sum_{1}^{m} {}_n \left\| \frac{u_m^n - u_m^{n-1}}{k} \right\|^2_{L^2(D)} \leqq C; \tag{2.22}$$

$$\max_{n=0,\cdots,m} \|u_m^n\|_V \leqq C \tag{2.23}$$

(throughout this paper we shall denote positive constants independent of m by C and C_i's). Note that by (2.16)

$$\begin{cases} w_m^n - w_m^{n-1} & \text{only if } u_m^n \geqq \rho_2 \\ w_m^n < w_m^{n-1} & \text{only if } u_m^n \leqq \rho_1 \quad \text{a.e. in } D, \text{ for } n = 1, \cdots, m. \\ |w_m^n - w_m^{n-1}| \leqq 2, \end{cases} \tag{2.24}$$

This yields

$$\sum_{1}^{m} {}_n |w_m^n - w_m^{n-1}| \leqq \frac{2}{\rho_2 - \rho_1} \sum_{1}^{m} {}_n |u_m^n - u_m^{n-1}| + 2 \quad \text{a.e. in } D, \tag{2.25}$$

whence

$$\left\| \sum_{1}^{m} {}_n |w_m^n - w_m^{n-1}| \right\|_{L^2(D)} \leqq \frac{2}{\rho_2 - \rho_1} \left\| \sum_{1}^{m} {}_n |u_m^n - u_m^{n-1}| \right\|_{L^2(D)} + 2(\text{meas } D)^{1/2}. \tag{2.26}$$

Let $u_m(x,t)$ denote the function obtained by linear interpolation of the values $u_m(x,nk)=u_m^n(x)$ for $n=0,\cdots,m$ a.e. in D; define w_m similarly. Set $\hat{u}_m(x,t)=u_m^n(x)$, $\hat{w}_m(x,t)=w_m^n(x)$ a.e. in D and $\hat{f}_m(t)=f_m^n$ in V' if $(n-1)k<t\leqq nk$, for $n=1,\cdots,m$. (2.15) becomes

$$u_{mt}+A\hat{u}_m+\hat{w}_m=\hat{f}_m \quad \text{in } V', \text{ a.e. in }]0,T[. \tag{2.27}$$

(2.22), (2.23) and (2.26) yield

$$\|u_m\|_{H^1(0,T;L^2(D))\cap L^\infty(0,T;V)}\leqq C, \tag{2.28}$$

(2.29)

$$\|w_m\|_{L^2(D;BV(0,T))}=\|w_m\|_{L^2(D;W^{1,1}(0,T))}\leqq\frac{2}{\rho_2-\rho_1}\|u_m\|_{L^2(D;W^{1,1}(0,T))}+2(\operatorname{meas}D)^{1/2}$$

$$\leqq\frac{2\sqrt{T}}{\rho_2-\rho_1}\|u_m\|_{H^1(0,T;L^2(D))}+2(\operatorname{meas}D)^{1/2}\leqq C,$$

moreover of course

$$\|w_m\|_{L^\infty(Q)}\leqq C,\ \|\hat{w}_m\|_{L^\infty(Q)}\leqq C. \tag{2.30}$$

(iii) *Limit.* By the above estimates there exist u, w such that, possibly taking subsequences,

$$u_m\to u \text{ in } H^1(0,T;L^2(D)) \text{ weak, in } L^\infty(0,T;V) \text{ weak star;} \tag{2.31}$$

$$\hat{u}_m\to 0 \text{ in } H^r(0,T;L^2(D)) \text{ weak } \forall r<\tfrac{1}{2}, \text{ in } L^\infty(0,T;V) \text{ weak star;} \tag{2.32}$$

$$w_m\to w \text{ in } L^2(D;BV(0,T)) \text{ weak star;} \tag{2.33}$$

$$w_m\to w,\ \hat{w}_m\to w \text{ in } L^\infty(Q) \text{ weak star.} \tag{2.34}$$

Taking $m\to\infty$ in (2.27) we get (2.6). As α and β are Lipschitz continuous we have

$$\alpha(\hat{u}_m)\to\alpha(u) \text{ in } H^r(0,T;L^2(D)) \text{ weak } \forall r<\tfrac{1}{2}, \text{ in } L^\infty(0,T;V) \text{ weak star;} \tag{2.35}$$

$$\beta(u_m)\to\beta(u) \text{ in } H^1(0,T;L^2(D)) \text{ weak, in } L^\infty(0,T;V) \text{ weak star.} \tag{2.36}$$

By (2.16)

$$w_m^n\in S(\alpha(u_m^n)) \quad \text{a.e. in } D, \text{ for } n=1,\cdots,m, \tag{2.37}$$

that is,

$$\hat{w}_m\in S(\alpha(\hat{u}_m)) \quad \text{a.e. in } Q. \tag{2.38}$$

Hence $\forall v\in L^1(Q)$

$$\iint_Q \hat{w}_m[\alpha(u_m)-v]\,dx\,dt\geqq\iint_Q\left(|\alpha(\hat{u}_m)|-|v|\right)dx\,dt, \tag{2.39}$$

whence taking $m\to\infty$ and using (2.34), (2.35)

$$\iint_Q w[\alpha(u)-v]\,dx\,dt\geqq\iint_Q\left(|\alpha(u)|-|v|\right)dx\,dt, \tag{2.40}$$

that is, (2.3).

Notice that (2.16) yields

$$w_m^n - w_m^{n-1} \in R\big(\beta(u_m^n)\big) \quad \text{a.e. in } D, \text{ for } n=1,\cdots,m. \tag{2.41}$$

Take a generic $v \in H^1(0,T,L^2(D))$ such that $\rho_1 \leqq v \leqq \rho_2$ a.e. in Q and $v(x,T) = \beta(u(x,T))$ a.e. in D; set $v_m^n(x) = v(x,nk)$ if $(n-1)k < t \leqq nk$ for $n=1,\cdots,m$ and denote the linear interpolate of these values by v_m. By (2.41)

$$0 \leqq \sum_1^m{}_n \int_D \big(w_m^n - w_m^{n-1}\big)\cdot\big(\beta(u_m^n) - v_m^n\big)\,dx = \iint_Q w_{mt}\cdot\big(\beta(\hat{u}_m) - \hat{v}_m\big)\,dx\,dt, \tag{2.42}$$

that is

$$0 \leqq \iint_D {}_{(C^0([0,T]))'}\langle w_{mt}, \beta(u_m) - v_m\rangle_{C^0([0,T])}\,dx + \iint_Q w_{mt}\cdot\big(\beta(u_m) - \beta(\hat{u}_m) + v_m - \hat{v}_m\big)\,dx\,dt. \tag{2.43}$$

The inclusion $H^1(Q) \subset L^2(D; C^0([0,T]))$ is compact, hence by (2.36) we have $\beta(u_m) \to \beta(u)$ in $L^2(D; C^0([0,T]))$ strong and $\beta(u_m) - \beta(\hat{u}_m) \to 0$ in $L^2(D; L^\infty(0,T))$ strong; the same convergence properties hold for v_m and $v_m - \hat{v}_m$; moreover by (2.33) $w_{mt} \to w_t$ in $L^2(D; (C^0([0,T]))')$ weak star and by (2.29) $\|w_{mt}\|_{L^2(D;\, L^1(0,T))} \leqq C$. Therefore taking $m \to \infty$ in (2.42) we get

$$0 \leqq \int_D {}_{(C^0([0,T]))'}\langle w_t, \beta(u) - v\rangle_{C^0([0,T])}\,dx + 0, \tag{2.44}$$

whence (2.4).

Finally $u \in C_S^0([0,T]; V)$ is entailed by the following result.

LEMMA 1. *Let X, Y be two Banach spaces, $X \subset Y$ with continuous injection, X being reflexive. Then*

$$L^\infty(0,T;X) \cap C_S^0([0,T]; Y) = C_S^0([0,T]; X). \tag{2.45}$$

Proof. Cf. [8, Chap. 3, §8.4]. □

PROPOSITION 1. *Assume that* (2.2), (2.8) *hold and that moreover*

$$f(0) + \Delta u^0 - w^0 \in L^2(D), \tag{2.46}$$

$$f = f_1 + f_2 \quad \text{with } f_1 \in W^{1,1}\big(0,T; L^2(D)\big), \qquad f_2 \in H^1(0,T; V'). \tag{2.47}$$

Then the problem (P1) *has at least one solution such that moreover*

$$u \in W^{1,\infty}\big(0,T; L^2(D)\big) \cap H^1(0,T; V), \tag{2.48}$$

$$w \in L^2\big(D; BV(0,T)\big). \tag{2.49}$$

Proof. It is sufficient to show stronger a priori estimates on the solutions u_m's of the approximated problems $(\mathrm{P1})_m$'s. $\forall m \in \mathbb{N}$, set

$$z_m^n = \frac{u_m^n - u_m^{n-1}}{k}, \quad \omega_m^n = \frac{w_m^n - w_m^{n-1}}{k}, \quad g_m^n = \frac{f_m^n - f_m^{n-1}}{k} \quad \text{for } n=1,\cdots,m;$$

set also $z_m^0 = f(0) + \Delta u^0 - w^0$ a.e. in D. Taking the incremental ratio w.r.t. time in (2.15), we get

$$\frac{z_m^n - z_m^{n-1}}{k} + A z_m^n + \omega_m^n = g_m^n \quad \text{in } V', \quad \text{for } n=1,\cdots,m. \tag{2.50}$$

Take a generic $l\in\{1,\cdots,m\}$, multiply (2.50) by kz_m^n and sum for $n=1,\cdots,l$. Notice that, as $w_m^n\in K(u_m^n,w_m^{n-1})$ and by the monotonicity of K. w.r.t. its first argument, we have

$$\omega_m^n z_m^n=\frac{w_m^n-w_m^{n-1}}{k}\cdot\frac{u_m^n-u_m^{n-1}}{k}\geqq 0 \quad \text{a.e. in } D, \text{ for } n=1,\cdots,m. \tag{2.51}$$

Thus by developments analogous to (2.17)–(2.21) and still by Gronwall's lemma we get

$$\max_{n=0,\cdots,m}\|z_m^n\|_{L^2(D)}\leqq C; \tag{2.52}$$

$$k\sum_1^m\|z_m^n\|_V^2\leqq C, \tag{2.53}$$

i.e.

$$\|u_m\|_{W^{1,\infty}(0,T;L^2(D))\cap H^1(0,T;V)}\leqq C. \tag{2.54}$$ □

2) We are going to introduce a weak formulation of (1.3), (1.18).

Assume that $\Gamma\equiv\partial D$ is a variety piecewise of class C^1, that D is only on one side of Γ. Set $W=H^1(D)$, Hilbert space with norm $\|v\|_W=[\int_D(v^2+|\overline{\nabla}v|^2)dx]^{1/2}$. Introduce the trace operator γ: $W\to H^{1/2}(\Gamma)$ (cf. [8, Chap. 1]) and, after identification of $L^2(\Gamma)$ with its dual, denote its adjoint by γ^*:

$$\begin{aligned}&\gamma^*\colon H^{-1/2}(\Gamma)\to W' \quad \forall v\in H^{-1/2}(\Gamma),\ \forall z\in W,\\ &{}_{W'}\langle\gamma^*v,z\rangle_W={}_{H^{-1/2}(\Gamma)}\langle v,\gamma z\rangle_{H^{1/2}(\Gamma)};\end{aligned} \tag{2.55}$$

in particular

$${}_{W'}\langle\gamma^*v,z\rangle_W=\int_\Gamma v\cdot\gamma z\,d\sigma \quad \forall v\in L^2(\Gamma),\ \forall z\in W. \tag{2.56}$$

Let (2.1) hold and

(2.57) $u^0\in L^2(D)\cap W^{1,1}(D)$ and $w^0\in L^\infty(\Gamma)$ such that $w^0\in S(\alpha(\gamma u^0))$ a.e. on Γ.

(P2) Find $u\in L^2(0,T;W)\cap H^1(0,T;W')$ $(\subset C^0([0,T];L^2(D)))$ and $w\in L^\infty(\Sigma)$ such that $\beta(\gamma u(\sigma,\cdot))\in C^0([0,T])$, $w(\sigma,\cdot)\in BV(0,T)$ a.e. on Γ and

(2.58) $w\in S(\alpha(\gamma u))$ a.e. on Σ,

(2.59) ${}_{(C^0([0,T]))'}\langle w_t,\beta(\gamma u)-v\rangle_{C^0([0,T])}\geqq 0$

$\forall v\in C^0([0,T])$ such that $\rho_1\leqq v\leqq\rho_2$, a.e. on Γ;

(2.60) $\lim_{t\to 0^+}[w(\sigma,t)-w^0(\sigma)]\cdot[\beta(u^0(\sigma))-v]\geqq 0$ $\forall v\in[\rho_1,\rho_2]$ a.e. on Γ;

(2.61) $u_t+Au+\gamma^*w=f$ in W' a.e. in $]0,T[$;

(2.62) $u(\sigma,0)=u^0(\sigma)$ a.e. on Γ.

THEOREM 2. *Assume that* (2.8), (2.47), (2.57) *hold and that*

$$f(0)+Au^0+\gamma^*w^0\in L^2(D). \tag{2.63}$$

Then problem (P2) *has at least one solution such that moreover*

$$u\in W^{1,\infty}(0,T;L^2(D))\cap H^1(0,T;W), \tag{2.64}$$

$$w\in L^2(\Gamma,BV(0,T)). \tag{2.65}$$

Proof. $\forall m \in \mathbb{N}$ we introduce to mean a time-discretized problem $(P2)_m$ similar to $(P1)_m$; also $(P2)_m$ has one (and only one) solution. Taking the incremental ration w.r.t. time in the approximate equation and using the procedure of Proposition 1, we get the following estimates

$$\|u_m\|_{W^{1,\infty}(0,T;L^2(D))\cap H^1(0,T;W)} \leqq C; \tag{2.66}$$

$$\|\hat{u}_m\|_{H^r(0,T;W)} \leqq C \quad \forall r < \tfrac{1}{2}. \tag{2.67}$$

The approximate equation has the form

$$u_{mt} + A\hat{u}_m + \gamma^* \hat{w}_m = \hat{f}_m \quad \text{in } W', \text{ a.e. in }]0,T[\tag{2.68}$$

(notations are as in the proof of Theorem 1). (2.25) yields

$$\begin{aligned} \|w_m\|_{L^2(\Gamma;BV(0,T))} &= \|w_m\|_{L^2(\Gamma;W^{1,1}(0,T))} \\ &\leqq \frac{2}{\rho_2 - \rho_1} \|\gamma u_m\|_{L^2(\Gamma;W^{1,1}(0,T))} + 2(\mathrm{meas}(\Gamma))^{1/2} \\ &\leqq C_1 \|u_m\|_{H^1(0,T;W)} + C_2 \leqq C. \end{aligned} \tag{2.69}$$

Therefore there exist u, w such that

$$u_m \to u \quad \text{in } W^{1,\infty}(0,T;L^2(D)) \text{ weak star, in } H^1(0,T;W) \text{ weak}; \tag{2.70}$$

$$\hat{u}_m \to u \quad \text{in } H^r(0,T;W) \text{ weak } \forall r < \tfrac{1}{2}; \tag{2.71}$$

$$w_m \to w \quad \text{in } L^2(\Gamma;BV(0,T)) \text{ weak star}; \tag{2.72}$$

$$w_m \to w,\ \hat{w}_m \to w \quad \text{in } L^\infty(Q) \text{ weak star}. \tag{2.73}$$

Taking $m \to \infty$ in (2.68) we get (2.61). (2.70) yields

$$\gamma u_m \to \gamma u \quad \text{in } H^1(0,T;H^{1/2}(\Gamma)) \text{ weak} \tag{2.74}$$

and as α and β are Lipschitz-continuous we get also

$$\alpha(\gamma \hat{u}_m) \to \alpha(\gamma u) \quad \text{in } H^r(0,T;L^2(\Gamma)) \cap L^2(0,T;H^{1/2}(\Gamma)) \text{ weak}, \ \forall r < \tfrac{1}{2}; \tag{2.75}$$

$$\beta(\gamma u_m) \to \beta(\gamma u) \quad \text{in } H^1(0,T;L^2(\Gamma)) \cap L^2(0,T;H^{1/2}(\Gamma)) \text{ weak}. \tag{2.76}$$

This allows us to prove (2.59) and (2.60) by a procedure analogous to that of the proof of Theorem 1. □

Remark. (2.63) is equivalent to the existence of a positive constant C such that

$$\left| {}_{V'}\langle f(0) + Au^0 + \gamma^* w^0, v \rangle_V \right| \leqq C \|v\|_{L^2(D)} \quad \forall v \in V; \tag{2.77}$$

thus if $f(0) \in V'$ is of the form

$${}_{V'}\langle f(0), v \rangle_V = \int_D \varphi \cdot v \, dx + {}_{H^{-1/2}(\Gamma)}\langle \psi, \psi v \rangle_{H^{1/2}}(\Gamma) \quad \forall v \in V \tag{2.78}$$

with $\varphi \in L^2(D)$ and $\psi \in H^{-1/2}(\Gamma)$, then (2.63) is equivalent to

$$\begin{aligned} &-\Delta u^0 + \varphi \in L^2(D) \quad \text{i.e. } \Delta u^0 \in L^2(D); \\ &\frac{\partial u^0}{\partial \nu} + w^0 + \psi = 0 \quad \text{in } H^{-1/2}(\Gamma). \end{aligned} \tag{2.79}$$

Remark. Theorems 1 and 2 can be extended to the case in which $D_t - \Delta$ is replaced by a nonlinear strictly parabolic operator.

3) We shall give a weak formulation of (1.3), (1.19). We introduce the interpolation spaces $H_{00}^{1/2}(D) = [L^2(D), V]_{1/2}$, $(H_{00}^{1/2}(D))' = [V', L^2(D)]_{1/2}$ as in [8, Chap. 1]. Assume that (2.1) and (2.2) hold with moreover

$$u^0 \in H_{00}^{1/2}(D); \tag{2.80}$$

let

$$u^1 \in \left(H_{00}^{1/2}(D)\right)'. \tag{2.81}$$

(P3) Find $u \in L^2(0,T;V) \cap H^2(0,T;V')$ and $w \in L^\infty(Q)$ such that $w(x,\cdot) \in BV(0,T)$ a.e. in D and

$$w \in S(\alpha(u)) \text{ a.e. in } Q; \tag{2.82}$$

$$_{(C^0([0,T]))'}\langle w_t, \beta(u) - v\rangle_{C^0([0,T])} \geqq 0 \tag{2.83}$$

$$\forall v \in C^0([0,T]) \text{ such that } \rho_1 \leqq v \leqq \rho_2, \qquad \text{a.e. in } D;$$

$$\lim_{t \to 0^+} \left[w(x,t) - w^0(x)\right] \cdot \left[\beta(u^0(x)) - v\right] \geqq 0 \quad \forall v \in [\rho_1, \rho_2], \text{ a.e. in } D; \tag{2.84}$$

$$u_{tt} + Au + w = f \quad \text{in } V', \text{ a.e. in }]0,T[; \tag{2.85}$$

$$u(x,0) = u^0(x) \quad \text{a.e. in } D; \tag{2.86}$$

$$u_t(x,0) = u^1(x) \quad \text{in } V'. \tag{2.87}$$

Remark. $u \in L^2(0,T;V) \cap H^2(0,T;V')$ entails $u \in H^1(0,T;L^2(D))$, hence $\beta(u(x,\cdot)) \in C^0([0,T])$ a.e. in D; moreover $L^2(0,T;V) \cap H^1(0,T;L^2(D)) \subset C^0([0,T]; H_{00}^{1/2}(D))$ and $H^1(0,T;L^2(D)) \cap H^2(0,T;V') \subset C^1([0,T]; (H_{00}^{1/2}(D))')$. Therefore (2.83), (2.86) and (2.87) are meaningful; furthermore (2.83) can be written equivalently in the form (1.15) a.e. in D, as $\beta(u) \in W^{1,1}(0,T)$ a.e. in D.

THEOREM 3. *Assume that* (2.1), (2.2) *hold and that*

$$u^0 \in V; \tag{2.88}$$

$$u^1 \in L^2(D). \tag{2.89}$$

Then problem (P3) *has at least one solution such that*

$$u \in C_S^1([0,T]; L^2(D)) \cap C_S^0([0,T]; V); \tag{2.90}$$

$$w \in L^2(D; BV(0,T)) \tag{2.91}$$

(*where* $C_2^1([0,T]; L^2(D)) = \{v \in C_S^0([0,T]; L^2(D)) \mid v_t \in C_S^0([0,T]; L^2(D))\}$).

Proof. Let $m \in \mathbb{N}$, $k = T/m$. Set

$$f_m^n = \frac{1}{k} \int_{(n-1)k}^{nk} f(t)\, dt \text{ in } V' \quad \text{for } n = 1, \cdots, m. \tag{2.92}$$

Remember the definition (2.14).

$(P3)_m$ Find $u_m^n \in V$, $w_m^n \in L^2(D)$ for $n = 1, \cdots, m$ such that, setting $u_m^0 = u^0$, $w_m^0 = w^0$, $z_m^0 = u^1$, $z_m^n = (u_m^n - u_m^{n-1})/k$ for $n = 1, \cdots, m$ a.e. in D,

$$\frac{z_m^n - z_m^{n-1}}{k} + Au_m^n + w_m^n = f_m^n \quad \text{in } V', \text{ for } n = 1, \cdots, m; \tag{2.93}$$

$$w_m^n \in K\left(u_m^n, w_m^{n-1}\right) \text{ a.e. in } D, \text{ for } n = 1, \cdots, m. \tag{2.94}$$

$\forall m \in \mathbb{N}$, $(P3)_m$ has one (and only one) solution, which can be constructed step by step as for $(P1)_m$.

Take a generic $l \in \{1, \cdots, m\}$, multiply (2.93) by $u_m^n - u_m^{n-1}$ and sum for $n = 1, \cdots, l$. Notice that

$$(2.95)\quad \sum_{n=1}^{l} \int_D \frac{z_m^n - z_m^{n-1}}{k} \cdot \left(u_m^n - u_m^{n-1}\right) dx$$

$$= \sum_{n=1}^{l} \int_D \left(z_m^n - z_m^{n-1}\right) z_m^n dx$$

$$\geqq \frac{1}{2} \sum_{n=1}^{l} \left(\left\| z_m^n \right\|_{L^2(D)}^2 - \left\| z_m^{n-1} \right\|_{L^2(D)}^2 \right) = \frac{1}{2} \left\| z_m^l \right\|_{L^2(D)}^2 - \frac{1}{2} \left\| u^1 \right\|_{L^2(D)}^2.$$

Thus by (2.17)–(2.21) we get (2.23) and

$$(2.96)\quad \max_{n=0,\cdots,m} \left\| \frac{u_m^n - u_m^{n-1}}{k} \right\|_{L^2(D)} \leqq C.$$

Therefore, using the notation of the proof of Theorem 1, we have (2.31)–(2.34) and

$$(2.97)\quad u_m \to u \quad \text{in } W^{1,\infty}\left(0, T; L^2(D)\right) \text{ weak star.}$$

The rest of the proof follows as in (2.35)–(2.44); just remark that by (2.85) $u_{tt} \in L^\infty(0, T; V')$, hence $u_t \in L^\infty(0, T; L^2(D)) \cap C^0([0, T]; V') \subset C_S^0([0, T]; L^2(D))$ by Lemma 1. □

3. Jump with anticipation.

1) Let (2.1) hold and

$$(3.1)\quad u^0 \in L^2(D).$$

We introduce a weak formulation of (1.17), (1.15).

(P4) Find $u \in L^2(0, T; V) \cap H^1(0, T; V')$ $(\subset C^0([0, T]; L^2(D)))$ and $w \in L^\infty(Q)$ such that $\beta(u) \in W^{1,1}(0, T; L^1(D))$ and

$$(3.2)\quad w \in S(\alpha(u)) \qquad \text{a.e. in } Q;$$

$$(3.3)\quad w \in S(\beta(u)_t) \qquad \text{a.e. in } Q;$$

$$(3.4)\quad u_t + Au + w = f \qquad \text{in } V', \text{ a.e. in }]0, T[,$$

$$(3.5)\quad u(x, 0) = u^0(x) \qquad \text{a.e. in } D.$$

THEOREM 4. *Assume that* (2.8) *holds and that*

$$(3.6)\quad f \in L^2(Q).$$

Then problem (P4) *has at least one solution such that moreover*

$$(3.7)\quad u \in H^1\left(0, T; L^2(D)\right) \cap C_S^0\left([0, T]; V\right) \cap L^2\left(0, T; H^2(D)\right).$$

Proof. Set

$$(3.8)\quad L(\xi, \eta) = S(\xi - \beta(\eta)) \quad \forall \xi, \eta \in \mathbb{R}.$$

Let $m \in \mathbb{N}$, $k = T/m$. Set

$$(3.9)\quad f_m^n(x) = \frac{1}{k} \int_{(n-1)k}^{nk} f(x, t)\, dt \quad \text{a.e. in } D, \text{ for } n = 1, \cdots, m.$$

$(P4)_m$ Find $u_m^n \in V$, $w_m^n \in L^\infty(D)$ for $n=1,\cdots,m$, such that, setting $u_m^0 = u^0$ a.e. in D,

$$\frac{u_m^n - u_m^{n-1}}{k} + Au_m^n + w_m^n = f_m^n \quad \text{in } V', \text{ for } n=1,\cdots,m; \tag{3.10}$$

$$w_m^n \in L\left(u_m^n, u_m^{n-1}\right) \quad \text{a.e. in } D, \text{ for } n=1,\cdots,m. \tag{3.11}$$

$\forall m \in \mathbb{N}$, $(P4)_m$ has one (and only one) solution, which can be constructed step by step as for $(P1)_m$. A priori estimates (2.22) and (2.23) can be obtained by the procedure (2.17)–(2.21); but here there is no reason for having (2.26). Using the notation introduced in the proof of Theorem 1, (3.10) can be written in the form

$$u_{mt} + A\hat{u}_m + \hat{w}_m = \hat{f}_m \quad \text{in } V', \text{ a.e. in }]0,T[; \tag{3.12}$$

as $\|\hat{f}_m\|_{L^2(Q)} \leqq C$, comparison in (3.12) yields $\|A\hat{u}_m\|_{L^2(Q)} \leqq C$, whence

$$\|\hat{u}_m\|_{L^2(0,T;H^2(D))} \leqq C. \tag{3.13}$$

Therefore there exist u, w, ξ such that, possibly taking subsequences

$$u_m \to u \quad \text{in } H^1(0,T;L^2(D)) \cap L^2(0,T;H^2(D)) \text{ weak}, \quad \text{in } L^\infty(0,T;V) \text{ weak star}; \tag{3.14}$$

$$\hat{u}_m \to u \quad \text{in } H^r(0,T;L^2(D)) \cap L^2(0,T;H^2(D)) \text{ weak } \forall r < \tfrac{1}{2}, \quad \text{in } L^\infty(0,T;V) \text{ weak star}; \tag{3.15}$$

$$w_m \to w \text{ and } \hat{w}_m \to w \quad \text{in } L^\infty(Q) \text{ weak star}; \tag{3.16}$$

$$u_m(T) \to \xi \quad \text{in } V \text{ weak}. \tag{3.17}$$

Thus also (2.35) and (2.36) hold; by (3.11) we have (2.37), whence (3.2) by the procedure (2.38)–(2.40). Let π denote the operator which extends any function defined in $[0,T]$ with value 0 in $\mathbb{R}\setminus[0,T]$; (3.12) can be written in the form

(3.18)

$$(\pi u_m)_t + \pi A\hat{u}_m + \pi\hat{w}_m = \pi\hat{f}_m + u_m(0)\delta_0(t) - u_m(T)\delta_0(t-T) \quad \text{in } (\mathcal{D}(0,T;V))'$$

(where δ_0 denotes the Dirac mass in 0); taking $m \to \infty$ we get

$$(\pi u)_t + Au + \pi w = f + u^0\delta_0(t) - \xi\delta_0(t-T) \quad \text{in } (\mathcal{D}(0,T;V))' \tag{3.19}$$

whence (3.4), (3.5) and

$$u(T) = \xi \quad \text{a.e. in } D. \tag{3.20}$$

Notice that, by the regularity of f, (3.4) can be written in the form

$$u_t - \Delta u + w = f \quad \text{a.e. in } Q. \tag{3.21}$$

Besides the notation introduced in the proof of Theorem 1, let $b_m(x,t)$ denote the function obtained by linear interpolation of the values $b_m(x,nk) = \beta(u_m^n(x))$ for $n = 0,\cdots,m$ a.e. in D. As β is Lipschitz-continuous, by (3.14) we have

$$b_m \to \beta(u) \quad \text{in } H^1(0,T;L^2(D)) \text{ weak, in } L^\infty(0,T;V) \text{ weak star}. \tag{3.22}$$

(3.11) yields

$$w_m^n \in S\left(\alpha\left(u_m^n\right)\right) \quad \text{for } n=1,\cdots,m; \tag{3.23}$$

$$w_m^n \in S\left(\beta\left(u_m^n\right) - \beta\left(u_m^{n-1}\right)\right) \quad \text{for } n=1,\cdots,m; \tag{3.24}$$

therefore

(3.25)

$$\sum_{n=1}^{m}\int_D w_m^n\big(u_m^n-u_m^{n-1}\big)\,dx$$

$$=\sum_{n=1}^{m}\int_D w_m^n\big[\beta(u_m^n)-\beta(u_m^{n-1})\big]\,dx+\sum_{n=1}^{m}\int_D w_m^n\big[\alpha(u_m^n)-\alpha(u_m^{n-1})\big]\,dx$$

$$\geqq\sum_{n=1}^{m}\|\beta(u_m^n)-\beta(u_m^{n-1})\|_{L^1(D)}+\sum_{n=1}^{m}\big(\|\alpha(u_m^n)\|_{L^1(D)}-\|\alpha(u_m^{n-1})\|_{L^1(D)}\big)$$

$$=k\sum_{n=1}^{m}\left\|\frac{\beta(u_m^n)-\beta(u_m^{n-1})}{k}\right\|_{L^1(D)}+\|\alpha(u_m^n)\|_{L^1(D)}-\|\alpha(u^0)\|_{L^1(D)};$$

thus multiplying (3.10) by $u_m^n-u_m^{n-1}$ and summing for $n=1,\cdots,m$, we get

$$\|u_{mt}\|^2_{L^2(Q)}+\frac{1}{2}\|\hat u_m(T)\|^2_V-\frac{1}{2}\|u^0\|^2_V+\|b_{mt}\|_{L^1(Q)}+\|\alpha(u_m(T))\|_{L^1(D)}$$
$$-\|\alpha(u^0)\|_{L^1(D)}\leqq\iint_Q\hat f_m\cdot u_{mt}\,dx\,dt,\tag{3.26}$$

whence taking $m\to\infty$ and using the lower semicontinuity of norms

$$\|u_t\|^2_{L^2(Q)}+\frac{1}{2}\|u(T)\|^2_V-\frac{1}{2}\|u^0\|^2_V+\|\beta(u)_t\|_{L^1(Q)}+\|\alpha(u(T))\|_{L^1(D)}$$
$$-\|\alpha(u^0)\|_{L^1(D)}\leqq\iint_Q f\cdot u_t\,dx\,dt.\tag{3.27}$$

Now multiply (3.4) by $u_t\in L^2(Q)$ and integrate w.r.t. space and time; using (3.2) we get

$$\|u_t\|^2_{L^2(Q)}+\frac{1}{2}\|u(T)\|^2_V-\frac{1}{2}\|u^0\|^2_V+\iint_Q w\cdot\beta(u)_t\,dx\,dt$$
$$+\|\alpha(u(T))\|_{L^1(D)}-\|\alpha(u^0)\|_{L^1(D)}=\iint_Q fu_t\,dx\,dt;\tag{3.28}$$

the comparison of the last two formulae yields

$$\iint_Q w\beta(u)_t\,dx\,dt\geqq\|\beta(u)_t\|_{L^1(D)}\tag{3.29}$$

whence, as $|w|\leqq1$ a.e. in Q,

$$\iint_Q w(\beta(u)_t-v)\,dx\,dt\geqq\iint_Q\big(|\beta(u)_t|-|v|\big)\,dx\,dt\quad\forall v\in L^1(Q)\tag{3.30}$$

that is (3.3). Finally by Lemma 1 we get $u\in C^0_S([0,T];V)$. □

PROPOSITION 2. *Assume that* (2.8) *and* (2.9) *hold. Then problem* (P4) *has at least one solution such that moreover*

$$u \in H^1(0,T;L^2(D)) \cap C_S^0([0,T];V). \tag{3.31}$$

Proof. As for Theorem 4, with the exception of (3.13); hence a priori u_m and $\hat{u}_m$ do not converge in $L^2(0,T;H^2(D))$ weak and the equation (3.4) does not hold a.e. in Q. Therefore in this case the multiplication of (3.4) by $u_t \notin L^2(0,T;V)$ is only formal. In order to make it rigorous, convolution with a regularizing kernel can be used, as in [7, Chap. 1, §1.8] (we shall give more details about this technique in the proof of Theorem 6). □

We give a regularity result.

PROPOSITION 3. *Assume that* (2.8), (2.46) *and* (2.47) *hold. Then* (P4) *has at least one solution such that moreover*

$$u \in W^{1,\infty}(0,T;L^2(D)) \cap H^1(0,T;V). \tag{3.32}$$

Proof. For the solutions of the approximate problems $(P3)_m$'s the stronger a priori estimate (2.54) can be proved as for Proposition 1 (notice that (2.51) holds also in this case). Thus we get in particular $u \in H^1(0,T;V)$; hence (3.4) can be multiplied by u_t.

2) We introduce another weak formulation of (1.17), (1.15). Let (2.8) and (2.9) hold.

(P4)′ Find $u \in H^1(0,T;L^2(D)) \cap L^\infty(0,T;V)$ $(\subset C_S^0([0,T];V))$ such that

$$\iint_Q \left[u_t(\alpha(u)-v)+\overline{\nabla} u\cdot\overline{\nabla}(\alpha(u)-v)\right]dx\,dt+\iint_Q \left(|\alpha(u)|-|v|\right)dx\,dt \leqq \int_0^T {}_{V'}\langle f,\alpha(u)-v\rangle_V dt \quad \forall v\in L^2(0,T;V), \tag{3.33}$$

$$\begin{aligned}&\iint_Q u_t(u_t-v)\,dx\,dt+\frac{1}{2}\int_D\left(|\overline{\nabla} u(T)|^2-|\nabla u^0|^2\right)dx-\iint_Q \overline{\nabla} u\cdot\overline{\nabla} v\,dx\,dt\\ &\quad+\iint_Q\left(|\beta(u)_t|-|v|\right)dx\,dt+\int_D\left(|\alpha(u(T))|-|\alpha(u^0)|\right)dx\\ &\leqq \iint_Q f_1(u_t-v)\,dx\,dt+{}_{V'}\langle f_2(T),u(T)\rangle_V-{}_{V'}\langle f_2(0),u^0\rangle_V\\ &\quad-\int_0^T {}_{V'}\langle f_{2t},u\rangle_V dt-\int_0^T\langle f_2,v\rangle_V dt \quad \forall v\in L^2(0,T;V).\end{aligned} \tag{3.34}$$

Remark. A system of variational inequalities was used also in [11] for the weak formulation of (1.3) coupled with (1.20).

PROPOSITION 4. *Assume that* (2.8) *and* (2.9) *hold. Then problem* (P4)′ *has at least one solution.*

Proof. Quite similar to that of Theorem 4. □

PROPOSITION 5. *Assume that* (2.8) *holds and that*

$$f \in L^2(Q) \cap W^{1,1}(0,T;V'). \tag{3.35}$$

Then any solution of problem (P4) *also solves* (P4)′.

Proof. (3.2) and (3.4) are equivalent to (3.33). As $f-w \in L^2(Q)$ then $u \in H^1(0,T;L^2(D)) \cap L^2(0,T;H^2(D))$ and (3.4) can be written in the form

$$u_t-\Delta u+w=f \quad \text{a.e. in } Q. \tag{3.36}$$

Multiply this by $u_t - v \in L^2(Q)$, for a generic $v \in L^2(0,T;V)$; notice that

$$(3.37) \qquad \iint_Q -\Delta u u_t \, dx\, dt = \iint_Q \overline{\nabla} u \cdot \overline{\nabla} u_t \, dx\, dt = \frac{1}{2} \int_D \left(|\overline{\nabla} u(T)|^2 - |\overline{\nabla} u^0|^2 \right) dx,$$

and that by (3.2), (3.3)

$$(3.38) \qquad \iint_Q w(u_t - v)\, dx\, dt = \iint_Q w\alpha(u)_t \, dx\, dt + \iint_Q w(\beta(u)_t - v)\, dx\, dt$$

$$\geqq \int_D \left(|\alpha(u(T))| - |\alpha(u^0)| \right) dx + \iint_Q \left(|\beta(u)_t| - |v| \right) dx\, dt;$$

thus we get (3.34). □

3) Assume that (2.1) and (3.1) hold. We introduce a weak formulation of (1.18), (1.15).

(P5) Find $u \in L^2(0,T;W) \cap H^1(0,T;W')$ $(\subset C^0([0,T];L^2(D)))$ and $w \in L^\infty(\Sigma)$ such that $\beta(\gamma u) \in W^{1,1}(0,T;L^1(\Gamma))$ and

$$(3.39) \qquad w \in S(\alpha(\omega)) \qquad \text{a.e. on } \Sigma;$$

$$(3.40) \qquad w \in S(\beta(\gamma u)_t) \qquad \text{a.e. on } \Sigma;$$

$$(3.41) \qquad u_t + Au + \gamma^* w = f \qquad \text{in } W', \text{ a.e. in }]0,T[;$$

$$(3.42) \qquad u(x,0) = u^0(x) \qquad \text{a.e. in } D.$$

Remark. It is possible to introduce a different setting, analogous to problem (P4)′, and to prove results similar to Propositions 4 and 5.

THEOREM 5. *Assume that* (2.8), (2.47) *and* (2.63) *hold. Then problem* (P5) *has at least one solution such that moreover*

$$(3.43) \qquad u \in W^{1,\infty}(0,T;L^2(D)) \cap H^1(0,T;W).$$

Proof. $\forall m \in \mathbb{N}$ we introduce a time-discretized problem $(P5)_m$ similar to $(P4)_m$; also $(P5)_m$ has one (and only one) solution, which can be constructed step by step. Taking the incremental ratio in the approximate equation and using a procedure similar to that of Proposition 1, we get the a priori estimates (2.66), (2.67) for the equation (2.68). Therefore there exist u, w such that possibly taking subsequences (2.70), (2.71) and (2.73) hold. Taking $m \to \infty$ in (2.68) we get (3.41). We have also (2.75) and (2.76). Therefore we can show (3.39) and (3.40) by a procedure similar to that used in the proof of Theorem 4 (just notice that the regularity $u \in H^1(0,T;W)$ is sufficient for multiplying (3.41) by u_t). □

4) Assume that (2.1), (2.80) and (2.81) hold. We introduce a weak formulation of (1.19), (1.15).

(P6) Find $u \in L^2(0,T;V) \cap H^2(0,T;V')$ and $w \in L^\infty(Q)$ such that

$$(3.44) \qquad w \in S(\alpha(u)) \qquad \text{a.e. in } Q;$$

$$(3.45) \qquad w \in S(\beta(u)_t) \qquad \text{a.e. in } Q;$$

$$(3.46) \qquad u_{tt} + Au + w = f \qquad \text{in } V', \text{ a.e. in }]0,T[;$$

$$(3.47) \qquad u(x,0) = u^0(x) \qquad \text{a.e. in } D;$$

$$(3.48) \qquad u_t(x,0) = u^1(x) \qquad \text{in } V'.$$

See the remark following (P3).

THEOREM 6. *Assume that* (2.1), (2.88) *and* (2.89) *hold. Then problem* (*P*6) *has at least one solution such that moreover*

$$u \in C_S^1([0,T];L^2(D)) \cap C_S^0([0,T];V). \tag{3.49}$$

Proof. $\forall m \in \mathbb{N}$ we introduce a time discretized problem $(P6)_m$ similar to $(P3)_m$, with (2.94) replaced by (3.11); also $(P6)_m$ has one (and only one) solution. Using the notations introduced in the proof of Theorem 1, the approximated equation can be written in the form

$$u_{mtt} + A\hat{u}_m + \hat{w}_m = \hat{f}_m \quad \text{in } V', \text{ for } n = 1, \cdots, m. \tag{3.50}$$

Multiplying the discretized equation by $u_m^n - u_m^{n-1}$, summing for $n = 1, \cdots, l$ and using (2.17)–(2.21) and (2.95) we get

$$\|u_m\|_{W^{1,\infty}(0,T;L^2(D)) \cap L^\infty(0,T;V)} \leqq C, \tag{3.51}$$

$$\|\hat{u}_m\|_{L^\infty(0,T;V)} \leqq C. \tag{3.52}$$

Fix a generic $\tilde{t} \in]0,t]$. By the above estimates there exist u, w, ξ, η such that possibly taking subsequences

$$u_m \to u \quad \text{in } W^{1,\infty}(0,T;L^2(D)) \cap L^\infty(0,T;V) \text{ weak star,} \tag{3.53}$$

$$\hat{u}_m \to u \quad \text{in } L^\infty(0,T;V) \text{ weak star;} \tag{3.54}$$

$$w_m \to w,\ \hat{w}_m \to w \quad \text{in } L^\infty(Q) \text{ weak star;} \tag{3.55}$$

$$u_m(\tilde{t}) \to \xi \quad \text{in } V \text{ weak;} \tag{3.56}$$

$$u_{mt}(\tilde{t}) \to \eta \quad \text{in } L^2(D) \text{ weak.} \tag{3.57}$$

By the procedure of (3.18)–(3.20), we get (3.46) and

$$\eta = u_t(\tilde{t}) \quad \text{a.e. in } D; \tag{3.58}$$

integrating (3.50) w.r.t. time and repeating the same procedure we obtain

$$\xi = u(\tilde{t}) \quad \text{a.e. in } D. \tag{3.59}$$

Multiplying (3.50) by $u_{mt} \in L^2(0,T;V)$ and integrating in $]0,\tilde{t}[$ we get (cf. (3.25), (3.26))

(3.60)

$$\frac{1}{2}\|u_{mt}(\tilde{t})\|_{L^2(D)}^2 - \frac{1}{2}\|u^1\|_{L^2(D)}^2 + \frac{1}{2}\|u_m(\tilde{t})\|_V^2 - \frac{1}{2}\|u^0\|_V^2 + \int_0^{\tilde{t}} \|b_{mt}\|_{L^1(D)} d\tau + \|\alpha(u_m(\tilde{t}))\|_{L^1(D)} - \|\alpha(u^0)\|_{L^1(D)} \leqq \int_0^{\tilde{t}} {}_{V'}\langle f_m, u_{mt}\rangle_V d\tau,$$

whence taking $m \to \infty$ and using lower semicontinuity of norms

(3.61)

$$\frac{1}{2}\|u_t(\tilde{t})\|_{L^2(D)}^2 - \frac{1}{2}\|u^1\|_{L^2(D)}^2 + \frac{1}{2}\|u(\tilde{t})\|_V^2 - \frac{1}{2}\|u^0\|_V^2 + \int_0^{\tilde{t}} \|\beta(u)_t\|_{L^1(D)} d\tau + \|\alpha(u(\tilde{t}))\|_{L^1(D)} - \|\alpha(0)\|_{L^1(D)} \leqq \int_0^{\tilde{t}} d\tau \int_D f_1 u_t dx + {}_{V'}\langle f_2(\tilde{t}), u(\tilde{t})\rangle_V - {}_{V'}\langle f_2(0), u^0\rangle_V - \int_0^{\tilde{t}} {}_{V'}\langle f_{2t}, u\rangle_V d\tau$$

and this last holds for any $\tilde{t}\in]0,T[$ (though the extracted subsequences may depend on $\tilde{t}$). Now we would like to multiply (3.46) by u_t and to integrate w.r.t. time (similarly to the procedure used in the proof of Theorem 4); but this is only formal, as $u_t\notin L^2(0,T;V)$ a priori. In order to make this rigorous, convolution with a regularizing kernel can be used; by the procedure of [7, Chap. 1, §1.8], one gets

$$\begin{aligned}(3.62)\quad &\frac{1}{2}\|u_t(\tilde{t})\|^2_{L^2(D)}-\frac{1}{2}\|u_t(t)\|^2_{L^2(D)}+\frac{1}{2}\|u(\tilde{t})\|^2_V-\frac{1}{2}\|u(t)\|^2_V\\ &\quad+\int_t^{\tilde{t}}d\tau\int_D w\beta(u)_t dx+\|\alpha(u(\tilde{t}))\|_{L^1(D)}-\|\alpha(u(t))\|_{L^1(D)}\\ &=\int_t^{\tilde{t}}d\tau\int_D f_1u_t dx+{}_{V'}\langle f_2(\tilde{t}),u(\tilde{t})\rangle_V-{}_{V'}\langle f_2(t),u(t)\rangle_V-\int_t^{\tilde{t}}{}_{V'}\langle f_{2t},u\rangle_V d\tau\\ &\qquad\qquad\text{a.e. for } t,\ \tilde{t}\in]0,T[.\end{aligned}$$

Now let $t\to 0$. By Lemma 1 $u_t\in C^0_S([0,T);L^2(D))$ and $u\in C^0_S([0,T];V)$, hence $u_t(t)\to u^1$ in $L^2(D)$ weak, $u(t)\to u^0$ in V weak; therefore by the lower semicontinuity of norms we get

$$\begin{aligned}(3.63)\quad &\frac{1}{2}\|u_t(\tilde{t})\|^2_{L^2(D)}-\frac{1}{2}\|u_1\|^2_{L^2(D)}+\frac{1}{2}\|u(\tilde{t})\|^2_V-\frac{1}{2}\|u^0\|^2_V\\ &\quad+\int_0^{\tilde{t}}d\tau\int_D w\beta(u)_t dx+\|\alpha(u(\tilde{t}))\|_{L^1(D)}-\|\alpha(u^0)\|_{L^1(D)}\\ &\geqq\int_0^{\tilde{t}}d\tau\int_D f_1u_t dx+{}_{V'}\langle f_2(\tilde{t}),u(\tilde{t})\rangle_V-{}_{V'}\langle f_2(0),u^0\rangle_V-\int_0^{\tilde{t}}{}_{V'}\langle f_{2t},u\rangle_V d\tau\\ &\qquad\qquad\text{a.e. for } \tilde{t}\in]0,T[;\end{aligned}$$

comparing (3.61) with (3.63) we get

$$(3.64)\qquad \int_0^{\tilde{t}}d\tau\int_D w\beta(u)_t dx\geqq\int_0^{\tilde{t}}\|\beta(u)_t\|_{L^1(D)}d\tau\quad\text{a.e. for } \tilde{t}\in]0,T[$$

whence (3.30), i.e. (3.45) □

Remark. As the system (1.15) is equivalent to (1.16), in (P4) we can replace (3.42), (3.44) by

$$(3.65)\qquad w\in S(\alpha(u)+\beta(u)_t)\quad\text{a.e. in } Q;$$

eliminating w by (3.4) and (3.65) we get

$$(3.66)\qquad u_t+Au+S(\alpha(u)+\beta(u)_t)\ni f\quad\text{in } V',\ \text{a.e. in }]0,T[.$$

Formally a solution u of (3.66) can be interpreted as a fixed-point for the application

$$(3.67)\quad \bar{u}\to\tilde{u}\quad\text{such that } \tilde{u}_t+A\tilde{u}+S(\tilde{u}-\beta(\bar{u})+\beta(\bar{u})_t)\ni f\quad\text{in } V',\ \text{a.e. in }]0,T[$$

as well as for

$$(3.68)\quad \bar{u}\to\tilde{u}\quad\text{such that } \tilde{u}_t+A\tilde{u}+S(\alpha(\bar{u})-\alpha(\bar{u})_t+\tilde{u}_t)\ni f\quad\text{in } V',\ \text{a.e. in }]0,T[$$

(where in both cases we have used the fact that $\alpha+\beta=$ Identity). (3.67) ((3.68)) corresponds to a variational inequality of the first type (of the second type, respectively). Therefore (3.66) can be regarded as a nonstandard nonstandard quasi-variational inequality.

4. Other results. Let ρ_1 and ρ_2 run along two sequences converging to the same limit, which we can assume to be zero: $\rho_{ij}\to 0$ as $j\to\infty$, for $i=1,2$; moreover let $\rho_{1j}\leqq\rho_{2j}\forall j$. Accordingly $\forall j\in\mathbb{N}$ define α_j and β_j similarly to (1.8) and define $(\tilde{\mathrm{P}}1)_j$ as (P1) with α and β replaced by α_j and β_j.

PROPOSITION 6. *Assume that* (2.1) *holds and that*

$$\forall j\in\mathbb{N}\quad w_j^0\in L^\infty(D)\ \text{and}\ w_j^0\in S\big(\alpha_j(u^0)\big)\quad a.e.\ \text{in}\ D; \tag{4.1}$$

$\forall j\in\mathbb{N}$ *let* u_j *be a solution of* $(\tilde{\mathrm{P}}1)_j$ (*existing by Theorem* 1). *Then*

$$u_j\to u\ \text{in}\ L^2(0,T;V)\cap H^1(0,T;V^1)\ \text{weak}, \tag{4.2}$$

where u *is the unique solution of the following variational inequality*:

Find $u\in L^2(0,T;V)\cap H^1(0,T;V')$ $(\subset C^0([0,T];L^2(D)))$ *such that*

$$\ {}_{V'}\langle u_t,u-v\rangle_V+\int_D\overline{\nabla}u\cdot\overline{\nabla}(u-v)\,dx+\int_D\big(|u|-|v|\big)\,dx\leqq{}_{V'}\langle f,u-v\rangle_V \tag{4.3}$$

$$\forall v\in V\ a.e.\ \text{in}\ ]0,T[;$$

$$u(x,0)=u^0(x)\quad a.e.\ \text{in}\ D. \tag{4.4}$$

Proof. $\forall j\in\mathbb{N}$, multiplying the corresponding (2.6) by u_j and integrating w.r.t. time, by a standard procedure (cf. (2.17)–(2.21)) we get

$$\|u_j\|_{L^\infty(0,T;L^2(D))\cap L^2(0,T;V)}\leqq\text{constant independent of } j \tag{4.5}$$

and then by comparison in (2.6)

$$\|u_j\|_{H^1(0,T;V')}\leqq\text{constant independent of } j. \tag{4.6}$$

Therefore there exists u such that, possibly taking a subsequence,

$$u_j\to u\quad\text{in}\ L^2(0,T;V)\cap H^1(0,T;V')\ \text{weak}; \tag{4.7}$$

hence

$$u_j\to u\quad\text{in}\ L^2(Q)\ \text{strong}, \tag{4.8}$$

whence, since $\lim_{j\to\infty}\alpha_j=$ identity uniformly in $\mathbb{R}$,

$$\alpha_j(u_j)\to u\quad\text{in}\ L^2(Q)\ \text{strong}. \tag{4.9}$$

$\forall j\in\mathbb{N}$, multiplying the corresponding (2.6) by u_j-v for a generic $v\in V$ and integrating w.r.t. time we get

$$\begin{aligned}\frac{1}{2}\|u_j(T)\|^2_{L^2(D)}-\frac{1}{2}\|u^0\|^2_{L^2(D)}-\int_0^T{}_{V'}\langle u_{jt},v\rangle_V dt+\|u_j\|^2_{L^2(0,T;V)}\\ -\iint_Q\overline{\nabla}u_j\cdot\overline{\nabla}v\,dx+\iint_Q w_j(u_j-v)\,dx=\int_0^T{}_{V'}\langle f,u_j-v\rangle_V dt.\end{aligned} \tag{4.10}$$

Notice that by (2.4) and (4.9)

$$\begin{aligned}&\lim_{j\to\infty}\iint_Q w_j(u_j-v)\,dx\,dt\\ &\quad=\lim_{j\to\infty}\iint_Q w_j\big(\alpha_j(u_j)-v\big)\,dx\,dt+\lim_{j\to\infty}\iint_Q w_j\big(u_j-\alpha_j(u_j)\big)\,dx\,dt\\ &\quad\geqq\iint_Q\big(|u|-|v|\big)\,dx\,dt;\end{aligned} \tag{4.11}$$

therefore taking the inferior limit as $j\to\infty$ in (4.10) we get

$$(4.12)\quad \frac{1}{2}\|u(T)\|^2_{L^2(D)}-\frac{1}{2}\|u^0\|^2_{L^2(D)}-\int_0^T {}_{V'}\langle u_t,v\rangle_V dt+\|u\|^2_{L^2(0,T;V)}$$
$$-\int_Q \overline{\nabla} u\cdot\overline{\nabla} v\,dx\,dt+\iint_Q (|u|-|v|)\,dx\,dt\leqq\int_0^T {}_{V'}\langle f,u-v\rangle_V dt\quad \forall v\in V,$$

which is equivalent to (4.3) and (4.4). Uniqueness of the solution of (4.3), (4.4) entails the convergence of the whole sequence $\{u_j\}_{j\in\mathbb{N}}$. □

PROPOSITION 7. *Assume that* (2.1) *and* (2.2) *hold. If*

$$(4.13)\qquad f\leqq 0 \text{ in } \mathcal{D}'(Q),\quad \rho_2\geqq 0,\quad u^0\leqq\rho_2 \text{ a.e. in } D,$$

then for any solution of problems (P1), (P2), (P3), *and* (P4)

$$(4.14)\qquad u\leqq\rho_2 \text{ a.e. in } D.$$

Similarly if $f\geqq 0$ *in* $\mathcal{D}'(Q)$, $\rho_1\leqq 0$, $u^0\geqq\rho_1$ *a.e. in* D, *then* $u\geqq\rho_1$ *a.e. in* D.

Proof. For problem (P1) it is sufficient to multiply (2.6) against $(u-\rho_2)^+$ in the first case, by $-(u-\rho_1)^-$ in the second one, then to integrate in time. Proof is quite similar for (P2), (P3) and (P4). □

For all of the above problems, uniqueness is an open question. We are just able to prove the following result.

PROPOSITION 8. *Assume that* (2.1), (2.2), (2.8) *hold and that*

$$(4.15)\qquad f=0 \text{ a.e. in } Q,\ \rho_1\leqq 0\leqq\rho_2,\ \rho_1\leqq u^0\leqq\rho_2 \text{ a.e. in } Q.$$

Then problem (P4) *has at most one solution such that* $u\in H^1(0,T;V)$.

Proof. Let (u_1,w_1), (u_2,w_2) be two solutions of (P4). By Proposition 7 $\rho_1\leqq u_i\leqq\rho_2$ a.e. in Q, hence $\beta(u_i)=u_i$ a.e. in Q for $i=1,2$. Take the difference of (2.6) written for u_1 and u_2, multiply it by $(u_1-u_2)_t\in L^2(0,T;V)$ and integrate w.r.t. time. Notice that as $w_i\in S(\beta(u_i)_t)$ $(i=1,2)$,

$$(4.16)\quad \iint_Q (w_1-w_2)\cdot(u_1-u_2)_t\,dx\,dt=\iint_Q (w_1-w_2)\cdot(\beta(u_1)_t-\beta(u_2)_t)\,dx\,dt\geqq 0;$$

thus by standard calculations we get $(u_1-u_2)_t=0$ a.e. in Q. □

Remark. All of the previous developments for problems (P1)–(P6) extend in a natural way to the case in which the normalized output w is replaced by $a(x,t)w$ with $a\in L^\infty(Q)$ ($a\in L^\infty(\Sigma)$ in (P2), (P5)), $a>0$ a.e. in Q (on Σ, respectively).

5. Stefan problem with hysteresis in the source. We remind that we denote by H the Heaviside graph, by λ the latent heat of water; if u represents temperature then $u+\lambda\chi\in u+\lambda H(u)$ has the physical meaning of enthalpy. We denote the inverse of $\xi\mapsto\xi+\lambda H(\xi)$ by l: $l(\xi)=(\xi-\lambda)^+-\xi^-\ \forall\xi\in\mathbb{R}$. Assume that (2.1) holds and let

$$(5.1)\quad \zeta^0\in\big(H^{1/2}_{00}(D)\big)'\cap L^1(D),\ w^0\in L^\infty(D) \text{ such that } w^0\in S\big(\alpha\big(l(\zeta^0)\big)\big) \text{ a.e. in } D$$

(ζ^0 and $l(\zeta^0)$ will represent initial enthalpy and initial temperature, respectively). We introduce a weak formulation of a two-phase Stefan problem with hysteresis in the source (see (1.7), (1.9), (1.20)):

(P7). Find $u\in L^2(0,T;V)$, $\chi\in L^\infty(Q)$, $w\in L^\infty(Q)$ such that $u+\lambda\chi\in H^1(0,T;V')$, $\beta(u(x,\cdot))\in C^0([0,T])$, $w(x,\cdot)\in BV(0,T)$ a.e. in D and

$$w\in S(\alpha(u)) \quad \text{a.e. in } Q, \tag{5.2}$$

$$_{(C^0[(0,T)])'}\langle w_t,\beta(u)-v\rangle_{C^0[(0,T)]}\geqq 0 \quad \forall v\in C^0[(0,T)] \text{ such that } \rho_1\leqq v\leqq\rho_2, \text{ a.e. in } D, \tag{5.3}$$

$$\lim_{t\to 0^+}\left[w(x,t)-w^0(x)\right]\cdot\left[\beta(u^0(x))-v\right]\geqq 0 \quad \forall v\in[\rho_1,\rho_2], \text{ a.e. in } D \tag{5.4}$$

$$\chi\in H(u) \quad \text{a.e. in } Q; \tag{5.5}$$

$$(u+\lambda\chi)_t+Au+w=f \quad \text{in } V', \text{ a.e. in }]0,T[; \tag{5.6}$$

$$(u+\lambda\chi)|_{t=0}=\zeta^0 \quad \text{in } V'. \tag{5.7}$$

Remark. $u+\lambda\chi\in L^2(Q)\cap H^1(0,T;V')\subset C^0([0,T];\ (H_{00}^{1/2}(D))')$, hence (5.7) is meaningful.

THEOREM 7. *Assume that* (2.8), (2.9) *and* (5.1) *hold. Then problem* (P7) *has at least one solution such that moreover*

$$u\in H^1(0,T;L^2(D))\cap L^\infty(0,T;V); \tag{5.8}$$

$$w\in L^2(D;BV(0,T)). \tag{5.9}$$

Proof. Let $m\in\mathbb{N}$, $k=T/m$. Define f_m^n and K as in (2.13), (2.14).

$(P7)_m$ Find $u_m^n\in V$, $\chi_m^n\in L^\infty(D)$, $w_m^n\in L^\infty(D)$ such that, setting $u_m^0=l(\zeta^0)$, $\chi_m^0=\zeta^0-l(\zeta^0)$, $w_m^0=w^0$ in D,

$$\frac{u_m^n-u_m^{n-1}}{k}+\frac{\chi_m^n-\chi_m^{n-1}}{k}+Au_m^n+w_m^n=f_m^n \quad \text{n } V', \text{ for } n=1,\cdots,m, \tag{5.10}$$

$$\chi_m^n\in H(u_m^n) \quad \text{a.e. in } D, \text{ for } n=1,\cdots,m; \tag{5.11}$$

$$w_m^n\in K(u_m^n,w_m^{n-1}) \quad \text{a.e. in } D, \text{ for } n=1,\cdots,m. \tag{5.12}$$

$\forall m\in\mathbb{N}$ $(P7)_m$ has one (and only one) solution, which can be constructed step by step as for $(P1)_m$; indeed, by the monotonicity of H, also in this case at every step a coercive, strictly convex, lower semicontinuous functional $V\to\mathbb{R}$ is to be minimized.

Multiply (5.10) by $u_m^n-u_m^{n-1}$ and sum for $n=1,\cdots,l$, for a generic $l\in\{1,\cdots,m\}$; notice that the monotonicity of H yields

$$\sum_{n=1}^{l}\int_D\frac{\chi_m^n-\chi_m^{n-1}}{k}\left(u_m^n-u_m^{n-1}\right)dx\geqq 0; \tag{5.13}$$

thus by (2.17)–(2.21) we get (2.22), (2.23).

We use the notation of the proof of Theorem 1. Moreover by χ_m we denote the function obtained by linear interpolation of the values $\chi_m^n(x,nk)=\chi_m^n(x)$ for $n=0,\cdots,m$, a.e. in D; set also $\hat\chi_m(x,t)=\chi_m^n(x)$ a.e. in D if $(n-1)k<t\leqq nk$ for $n=1,\cdots,m$. (5.10) can be written in the form

$$u_{mt}+\lambda\chi_{mt}+A\hat u_m+\hat w_m=\hat f_m \text{ in } V', \quad \text{a.e. in }]0,T[; \tag{5.14}$$

we have a priori estimates (2.28)–(2.30). Therefore there exist u, χ, w such that possibly by taking subsequences (2.31)–(2.34) hold and moreover

$$\chi_m\to\chi \quad \text{in } L^\infty(Q) \text{ weak star.} \tag{5.15}$$

We have $\hat{\chi}_m \in H(\hat{u}_m)$ a.e. in Q; hence by (2.31) and by a standard monotonicity technique we get (5.5). The rest of the proof follows as for Theorem 1. □

It does not seem trivial to extend Theorems 2, 4, 5 to the case in which the linear operator $u \mapsto u_t + Au$ is replaced by $u \mapsto u_t + \lambda\chi_t + Au$ where $\chi \in H(u)$, as in the Stefan problem.

A result analogous to Proposition 6 holds for (P7), too.

6. A parabolic system with hysteresis. Hoppensteadt and Jäger have modelled and numerically studied a biological phenomenon exhibiting hysteresis (cf. [3], [4]). They formulated the following parabolic system

$$u_{it} - D_i \Delta u_i + c_i s = 0 \qquad (i = 1,2), \tag{6.1}$$

where D_i and c_i are positive constants, $u_i \geqq 0$ $(i = 1,2)$; s is related to (u_1, u_2) as follows: the quadrant $(\mathbb{R}^+)^2$ is parted into three sets by two disjoint curves on which s is switched from 0 to 1 and conversely; let $\Gamma_1(\Gamma_2)$ correspond to the switching off (on) curve (see Fig. 3). We can assume Γ_1 and Γ_2 to be level curves of a "smooth" function φ: $(\mathbb{R}^+)^2 \to \mathbb{R}$; that is $\Gamma_i = \{(u_1, u_2) \mid \varphi(u_1, u_2) = \rho_i\}$ $(i = 1,2)$ with ρ_1, $\rho_2 \in \mathbb{R}$. If $\rho_1 < \rho_2$, then the relation between $2s - 1$ and $\varphi(u_1, u_2)$ is the same as that between w and u in (1.2). Therefore the system (6.1) can be studied similarly to the single equation (1.17).

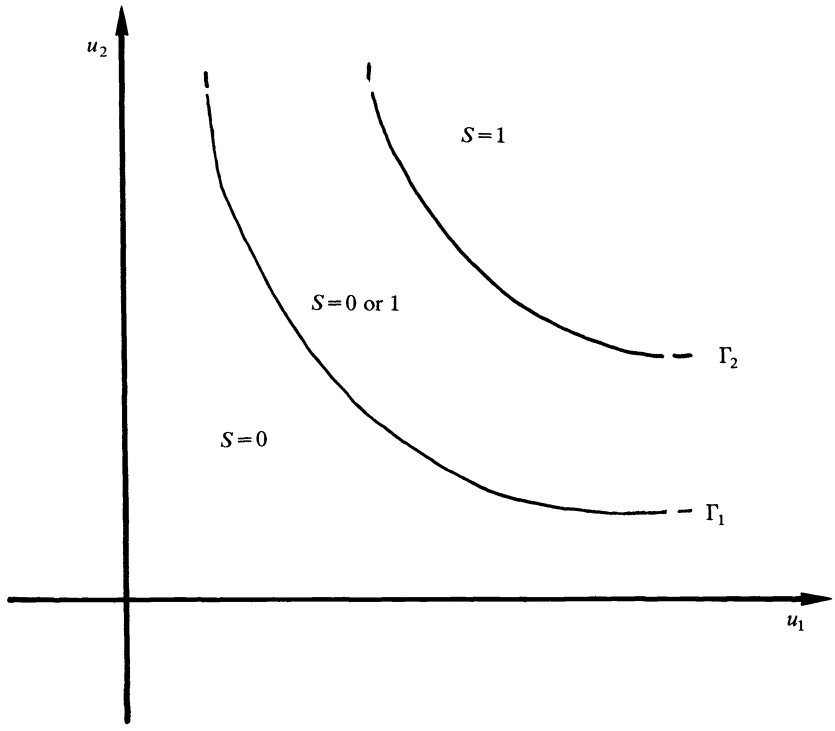


FIG. 3

In the biological phenomenon considered by Hoppensteadt and Jäger, the u_i's represent the concentrations of substances ("nutrients") which activate the growth of another substance ("bacteria"). This last diffuses slowly and reacts fast, so that one can use the ordinary differential equation $b_t = c_0 s$, where b represents the concentration of bacteria and c_0 is a positive constant. Due to the threshold effect in the dependence of

s on (u_1, u_2), some spatial structures appear in the distribution of b. This is analogous to the Liesegang phenomenon arising in chemistry, in which nutrients and bacteria are replaced by ions. The above model has also been confirmed by numerical tests (cf. [3], [4]).

Let $\tilde{V}=(H^1(\Omega))^2$, $\tilde{H}=(L^2(D))^2$; identifying $\tilde{H}$ with its dual $\tilde{H}'$, we get the Hilbert triple $\tilde{V}\subset\tilde{H}=\tilde{H}'\subset\tilde{V}'$, with dense and compact inclusions. Let A: $\tilde{V}\to\tilde{V}'$ be linear, continuous, symmetric and coercive (so that $u\mapsto({}_{\tilde{V}'}\langle Au,u\rangle_{\tilde{V}})^{1/2}$ defines a norm equivalent to $\|u\|_{\tilde{V}}$); let $f\in L^2(0,T;\tilde{V}')$ and $u^0\in\tilde{V}'$. Moreover let $c=(c_1,c_2)\in(\mathbb{R}^+)^2$ and φ: $(\mathbb{R}^+)^2\to\mathbb{R}$ continuous.

(P8) Find $u=(u_1,u_2)\in L^2(0,T;\tilde{V})\cap H^1(0,T;\tilde{V}')$ and $w\in L^\infty(Q)$ such that, setting $z=\varphi(u)$ a.e. in Q,

$$\beta(z(x),\cdot)\in C^0([0,T]),\ w(x,\cdot)\in BV(0,T)\quad\text{a.e. in } D; \tag{6.2}$$

$$w\in S(\alpha(z))\quad\text{a.e. in } Q; \tag{6.3}$$

$$({}_{C^0([0,T])'}\langle w_t,\beta(z)-v\rangle_{C^0([0,T])}\geqq 0 \qquad \forall v\in C^0([0,T])\text{ such that }\rho_1\leqq v\leqq\rho_2,\text{ a.e. in } D; \tag{6.4}$$

$$\lim_{t\to 0^+}\left[w(x,t)-w^0(x)\right]\cdot\left[\beta(u^0(x))-v\right]\geqq 0\quad\forall v\in[\rho_1,\rho_2],\quad\text{a.e. in } D \tag{6.5}$$

$$u_t+Au+c\frac{w+1}{2}=f\quad\text{in }\tilde{V}',\text{ a.e. in }]0,T[; \tag{6.6}$$

$$u(x,0)=u^0(x)\quad\text{a.e. in } D. \tag{6.7}$$

Remark that (6.6) corresponds to (2.6) setting

$$ {}_{\tilde{V}'}\langle Au,v\rangle_{\tilde{V}}=\int_D\left(D_1\overline{\nabla}u_1\cdot\overline{\nabla}v_1+D_2\overline{\nabla}u_2\cdot\overline{\nabla}v_2\right)dx,$$

$$ {}_{\tilde{V}'}\langle f,v\rangle_{\tilde{V}}=\int_\Gamma\left(D_1\frac{\partial u_1}{\partial\nu}\cdot v_1+D_2\frac{\partial u_2}{\partial\nu}\cdot v_2\right)d\sigma$$

a.e. in $]0,T[$ $\forall v=(v_1,v_2)\in V$ and $s=(w+1)/2$.

THEOREM 8. *Assume that*

$$\varphi\text{ is Lipschitz-continuous and monotone with respect to each of its arguments}; \tag{6.8}$$

$$u^0\in\tilde{V}; \tag{6.9}$$

$$f=f_1+f_2,\quad f_1\in L^2(0,T;\tilde{H}),\quad f_2\in W^{1,1}(0,T;\tilde{V}). \tag{6.10}$$

Then problem (P8) *has at least one solution such that moreover*

$$u\in H^1(0,T;\tilde{H})\cap C_S^0([0,T];\tilde{V}), \tag{6.11}$$

$$w\in L^2(D;BV(0,T)). \tag{6.12}$$

Proof. Similar to that of Theorem 1. An approximated problem $(P8)_m$ analogous to $(P1)_m$ can be introduced, with (2.16) replaced by

$$w_m^n\in K\left(\varphi(u_m^n),w_m^{n-1}\right)\quad\text{a.e. in } D,\text{ for } n=1,\cdots,m. \tag{6.13}$$

By (6.8) $K(\varphi(\cdot),w_m^{n-1}(x))$ is a maximal monotone operator a.e. in D; therefore also $(P8)_m$ has one (and only one) solution, which can be constructed step by step. A priori estimates (2.22), (2.23) can be proved also here (obviously with $L^2(D)$ and V replaced by $\tilde{H}$ and $\tilde{V}$, respectively).

We use the notation introduced in the proof of Theorem 1. Moreover by $z_m(x,t)$ we denote the function obtained by linear interpolation of the values $z_m(x,nk)=\varphi(u_m^n(x))$ for $n=0,\cdots,m$, a.e. in D; set $\hat{z}_m(x,t)=\varphi(u_m^m(x))$ a.e. in D if $(n-1)k<t\leqq nk$ for $n=1,\cdots,m$. Thus the approached equation can be written in the form (2.27) and also (2.28) holds. Then by (6.8) we have

$$\|z_m\|_{H^1(0,T;\tilde{H})\cap L^\infty(0,T;\tilde{V})}\leqq C \tag{6.14}$$

whence as in (2.29)

$$\|w_m\|_{L^2(D;BV(0,T))}\leqq C. \tag{6.15}$$

Therefore there exist u, z, w such that, possibly taking subsequences, (2.31)–(2.34) hold and

$$z_m\to z \text{ in } H^1(0,T;\tilde{H})\cap L^\infty(0,T;\tilde{V}) \text{ weak star.} \tag{6.16}$$

This last and (2.31) yield

$$z=\varphi(u) \quad \text{a.e. in } D. \tag{6.17}$$

The rest of the proof follows as in (2.35)–(2.44). □

Acknowledgment. The author is indebted to the referees for several useful suggestions.

REFERENCES

[1] K. GLASHOFF AND J. SPREKELS, *An application of Glicksberg's theorem to set-valued integral equations arising in the theory of thermostats*, this Journal, 12 (1981), pp. 477–486.

[2] _____, *The regulation of temperature by thermostats and set-valued integral equations*, J. Integral Equations, 4 (1982), pp. 95–112.

[3] F. C. HOPPENSTEADT AND W. JÄGER, *Pattern formation by bacteria*, in Lecture Notes in Biomathematics 38, Springer, Berlin, 1980, pp. 69–81.

[4] W. JÄGER, *A diffusion reaction system modelling spatial patterns*, Equadiff, Bratislava 1981, Teubner Texte zur Math., 47, 1982.

[5] M. A. KRASNOSEL'SKIĬ, *Equations with non-linearities of hysteresis type*, (Russian), VII. Int.Konf. Nichtlineare Schwing., Berlin 1975; Abh. Akad. Wiss. DDR, Jahrg. 1977, 3 (1977), pp. 437–458. (English abstract in Zentralblatt für Mathematik 406-93032).

[6] M. A. KRASNOSEL'SKIĬ AND A. V. POKROVSKIĬ, *Operators representing non-linearities of hysteresis type*, in Theory of Operators in Functional Spaces, G. P. Akilov, ed., Nauka, Novosibirsk, 1977. (In Russian)

[7] J. L. LIONS, *Quelques méthodes de résolution des problèmes aux limites non linéaires*, Dunod, Gauthier-Villars, Paris, 1969.

[8] J. L. LIONS AND E. MAGENES, *Non Homogeneous Boundary Value Problems*, Vol. I Grund. Math. Wiss. 181, Springer, Berlin, 1972.

[9] P. A. POLUBARINOVA AND KOCHINA, *Theory of Ground Water Motion*, 2nd ed. Nauka, Moskwa 1977. (In Russian)

[10] A. VISINTIN, *A model for hysteresis of distributed systems*, Ann. Mat. Pura. Appl., 131 (1982), pp. 203–231.

[11] _____, *A phase transition problem with delay*, Control and Cybernetics, 1-2 (1982), pp. 5–18.

[12] _____, *On Preisach' model for hysteresis*, J. Nonlinear Analysis T. M. A., 9 (1984), pp. 977–996.

[13] _____, *Partial differential equations with hysteresis functionals*, Proc. 6th International Conference on Computing Methods in Applied Sciences and Engineering, Versailles, 1983.

SIAM J. MATH. ANAL.
Vol. 17, No. 5, September 1986

007

ON THE FRACTAL DIMENSION OF ATTRACTORS FOR VISCOUS INCOMPRESSIBLE FLUID FLOWS*

J.-M. GHIDAGLIA†

Abstract. We prove that attractors associated to various equations of viscous incompressible fluid flows have finite fractal dimension and lie in the set of C^∞-functions. The first part of this article consists in an abstract formulation of our results; then in the second part, we apply these results to Navier–Stokes equations with nonhomogeneous boundary conditions to N.S.E. on a Riemannian manifold, to thermo-hydraulic equations and finally to magnetohydrodynamic equations.

Key words. Navier–Stokes equations, long-time behavior, attractors, fractal dimension

AMS(MOS) subject classifications. Primary 35B40, 35Q10, 35K55, 58G11, 76F99, 76W05, 76D05

Introduction. The equations which describe the motion of a viscous fluid can be viewed as an infinite dynamical system. However from a physical point of view it is admitted that this motion depends on a finite number of degrees of freedom. This is a physical evidence for laminar flows (observed for small Reynolds numbers) which correspond mathematically to the existence of a stable steady state. For large Reynolds numbers, the phenomenon of Turbulence occurs and the interpretation becomes somewhat more complicated (both physically and mathematically).

D. Ruelle and F. Takens [28] attribute Turbulence to strange attractors. These sets represent the behaviour of the flow since after some transient period they trap the motion. Thus in a certain sense, attractors carry all the information contained in the external excitation. This feature is characteristic of dissipative systems.

Proving that attractors have finite dimension is a way to give a mathematical sense to the physical concept of finite number of degrees of freedom. In 1979, C. Foias and R. Temam [14] proved that attractors associated to Navier–Stokes equations have finite dimension and more recently M. Sermange and R. Temam [29] have investigated the magnetohydrodynamic equations. Using completely different methods, various authors have obtained bounds for the dimension of attractors which have physical interest (i.e. in terms of nondimensional physical numbers) in case of Navier–Stokes equations with homogeneous boundary condition: C. Foias and R. Temam [15], P. Constantin and C. Foias [8], P. Constantin, C. Foias and R. Temam [10], D. Ruelle [26], E. Lieb [23], P. Constantin, C. Foias, O. Manley and R. Temam [9], R. Temam [33] (this reference contains a quasi optimal result for 2D-N.S.E.) and in a special case of thermo-hydraulic equations, D. Ruelle [27].

In this article we shall study various situations arising in fluid mechanics of viscous incompressible fluid flows. We successively consider the motion of a fluid dragged by moving walls (Navier–Stokes equations with nonhomogeneous boundary conditions), the motion of a fluid on a Riemannian manifold (motivated by a meteorological model for the circulation of the atmosphere), the motion of a fluid submitted to buoyancy effects (thermo-hydraulic equations). Finally we consider the motion of a resistive fluid (magnetohydrodynamic equations).

*Received by the editors November 14, 1984, and in revised form April 5, 1985.
†Laboratoire d'Analyse Numérique, Université Paris-Sud, 91405 Orsay, France.

The plan is as follows. In the first part, we consider in §1.1, an abstract equation (generalizing the usual Navier–Stokes equation where $\mathfrak{R}=0$):

$$\frac{d\phi}{dt}+\mathfrak{A}\phi+\mathfrak{B}(\phi,\phi)+\mathfrak{R}\phi=\mathfrak{F}, \tag{0.1}$$

and briefly recall the well-known results of existence and uniqueness. Estimates, uniform with respect to time, are stated in §1.2. They extend those obtained by C. Guillopé [20] for the Navier–Stokes equations with homogeneous boundary conditions. In §1.3, which contains the main results of the first part, we introduce with C. Foias and R. Temam [14], R. Temam [31], a notion of functional invariant set and attractor. Then we prove that the attractors have finite fractal dimension and lie in a set of regular functions. In the second part we show how the results of the first part apply to the situations mentioned previously, namely Navier–Stokes equations (N.S.E.) with nonhomogeneous boundary conditions (§2.1), N.S.E. on a Riemannian manifold (§2.2), thermo-hydraulic equations (§2.3) and finally M. H. D. equations (§2.4).

In a subsequent work we shall derive bounds for the dimension of attractors in term of nondimensional physical numbers similar to those obtained by P. Constantin, C. Foias, O. Manley and R. Temam [9] in the case of the Navier–Stokes equations.

This work summarizes a thesis [16] prepared at the Laboratoire d'Analyse Numérique of the University of Paris-Sud at Orsay, France.

1. Review of known results and complements.

1.1. The abstract framework. Let $\mathfrak{v}$ be a separable Hilbert space topologically included in a Hilbert space $\mathfrak{H}$ with compact injection; we denote by $(\cdot,\cdot)$ and $|\cdot|$ the scalar product and the norm on $\mathfrak{H}$ and by $|\cdot|_{\mathfrak{v}}$ the norm on $\mathfrak{v}$. We suppose that $\mathfrak{v}$ is dense in $\mathfrak{H}$ and thus, identifying $\mathfrak{H}$ with its dual, we have the usual injections

$$\mathfrak{v}\subset\mathfrak{H}\subset\mathfrak{v}'. \tag{1.1}$$

We shall also denote by $(\cdot,\cdot)$ the pairing between $\mathfrak{v}$ and $\mathfrak{v}'$.

Let $\mathfrak{A}$ be a self-adjoint linear operator, continuous form $\mathfrak{v}$ into $\mathfrak{v}'$, such that $(\alpha>0)$:

$$(\mathfrak{A}\phi,\phi)\geqq\alpha|\phi|_{\mathfrak{v}}^2 \quad \forall\phi\in\mathfrak{v}. \tag{1.2}$$

For convenience we shall use in the sequel the norm $\|\phi\|\equiv(\mathfrak{A}\phi,\phi)^{1/2}$ on $\mathfrak{v}$. This norm is equivalent to $|\cdot|_{\mathfrak{v}}$.

Let $\mathfrak{R}$ be a linear continuous operator from $\mathfrak{v}$ into $\mathfrak{v}'$ which maps $D(\mathfrak{A})\equiv\{\phi\in\mathfrak{v},\ \mathfrak{A}\phi\in\mathfrak{H}\}$ into $\mathfrak{H}$ and such that there exist $\theta_1,\theta_2\in[0,1[$ and two positive constants K_1 and K_2 satisfying

$$|\mathfrak{R}\phi|\leqq K_1\|\phi\|^{1-\theta_1}|\mathfrak{A}\phi|^{\theta_1} \quad \forall\phi\in D(\mathfrak{A}); \tag{1.3}$$

$$|(\mathfrak{R}\phi,\phi)|\leqq K_2\|\phi\|^{1+\theta_2}|\phi|^{1-\theta_2} \quad \forall\phi\in\mathfrak{v}. \tag{1.4}$$

Finally let $\mathfrak{B}$ be a bilinear continuous operator from $\mathfrak{v}\times\mathfrak{v}$ into $\mathfrak{v}'$ and $D(\mathfrak{A})\times D(\mathfrak{A})$ into $\mathfrak{H}$ such that

$$(\mathfrak{B}(\phi,\psi),\psi)=0 \quad \forall\phi,\psi\in\mathfrak{v}; \tag{1.5}$$

and there exist $\theta_3, \theta_4 \in [0,1[$ and $K_3, K_4 > 0$ satisfying

$$|(\mathfrak{B}(\phi_1,\phi_2),\phi_3)| \leqq K_3 \|\phi_1\| \|\phi_2\| \|\phi_3\|^{\theta_3} |\phi_3|^{1-\theta_3} \quad \forall \phi_i \in \mathfrak{v}, \tag{1.6}$$

$$|\mathfrak{B}(\phi_1,\phi_2)| + |\mathfrak{B}(\phi_2,\phi_1)| \leqq K_4 \|\phi_1\| \|\phi_2\|^{1-\theta_4} |\mathfrak{A}\phi_2|^{\theta_4} \quad \forall \phi_1 \in \mathfrak{v} \quad \forall \phi_2 \in D(\mathfrak{A}). \tag{1.7}$$

Flow of strong solutions. Given $\phi^0 \in \mathfrak{v}$ and $\mathfrak{F} \in L^\infty(0,T;\mathfrak{H})$ with $T>0$, the initial value problem

$$\frac{d\phi}{dt} + \mathfrak{A}\phi + \mathfrak{B}(\phi,\phi) + \mathfrak{R}\phi = \mathfrak{F}, \tag{1.8}$$

$$\phi(0) = \phi^0, \tag{1.9}$$

has a unique solution on $[0, T_1(\|\phi^0\|)]$ $(0 < T_1 \leqq T)$ such that

$$\phi \in \mathfrak{C}([0,T_1];\mathfrak{v}) \cap L^2(0,T_1;D(\mathfrak{A})) \tag{1.10}$$

where $T_1 = T_1(\|\phi^0\|)$ depends on $\|\phi^0\|$ and the other data $\mathfrak{A}$, $\mathfrak{B}$, $\mathfrak{R}$, $\mathfrak{F}$; more precisely there exists a positive constant C_1 depending only on the operators $\mathfrak{A}$, $\mathfrak{B}$, $\mathfrak{R}$ and $\mathfrak{F}$ such that

$$T_1(\rho) \geqq \frac{C_1}{(1+\rho^2)^{1/(1-\theta_4)}}, \tag{1.11}$$

$$\sup_{0 \leqq t \leqq T_1} \|\phi(t)\| \leqq 1 + 2\|\phi^0\|. \tag{1.12}$$

DEFINITION 1.1. We say that ϕ is a strong solution of (1.1.8) on an interval I if $\phi \in C(I;V) \cap L^2_{\text{loc}}(I;D(\mathfrak{A}))$.

If $\mathfrak{B}$ satisfies in addition: there exist $\theta_5 \in [0,1[$, $K_5 > 0$, such that

$$|\mathfrak{B}(\phi_1,\phi_2)| \leqq K_5 \|\phi_1\|^{1-\theta_5} |\phi_1|^{\theta_5} \|\phi_2\|^{1-\theta_5} |\mathfrak{A}\phi_2|^{\theta_5} \quad \forall \phi_1 \in \mathfrak{v}, \quad \forall \phi_2 \in D(\mathfrak{A}); \tag{1.13}$$

we can take $T_1 = T$ i.e. we have existence for all time. These results are well known in the case of Navier–Stokes equations (C. Foias and G. Prodi [13], J. L. Lions [24], R. Temam [31, 32], $\cdots$); the details of the proofs, for the equation considered here, are given in [16].

1.2. Time uniform estimates. We denote by $\mathfrak{v}_m = D(\mathfrak{A}^{m/2})$ the scale of Hilbert spaces endowed with the norm $\|\phi\|_m \equiv |\mathfrak{A}^{m/2}\phi|$.

We introduce a family of Hilbert spaces $\{\mathfrak{E}_m\}_{m \in \mathbb{N}}$, with

$$\mathfrak{E}_{m+1} \subset \mathfrak{E}_m \quad \forall m \in \mathbb{N}, \tag{2.1}$$

the injection being continuous,

$$\begin{aligned} &\mathfrak{v}_m \text{ is a closed subspace of } \mathfrak{E}_m, \forall m \in \mathbb{N}, \text{ the norm} \\ &\text{induced by } \mathfrak{E}_m \text{ on } \mathfrak{v}_m \text{ being equivalent to } \|\cdot\|_m. \end{aligned} \tag{2.2}$$

We assume that $\mathfrak{E}_0 = \mathfrak{H}$ and

$$\mathfrak{A}^{-1} \text{ is continuous from } \mathfrak{E}_m \text{ into } \mathfrak{E}_{m+2} \cap \mathfrak{v} \quad \forall m \geqq 0. \tag{2.3}$$

We make the following assumptions on $\mathfrak{B}$ and $\mathfrak{R}$:

(2.4) $\mathfrak{R}$ is continuous from $\mathfrak{E}_{m+1}$ into $\mathfrak{E}_m$, $m \geqq 1$;

(2.5) $\mathfrak{B}$ is continuous from $\mathfrak{E}_{m+1} \times \mathfrak{E}_{m+1}$ into $\mathfrak{E}_m$, $m \geqq 1$.

The following theorem was first proved by C. Guillopé [20] for Navier–Stokes equations with homogeneous boundary conditions. The proof in our more general case can be found in [16].

Let $m \in \mathbb{N}$ and $l = [(m/2)]$.

THEOREM 2.1. *If* $\mathfrak{F}$ *satisfies*

(2.6) $$\mathfrak{F}^{(j)} \in \mathfrak{C}_b([0, +\infty[;\ \mathfrak{E}_{m-2j-2}), \qquad j = 0, \cdots, l-1,$$

(2.7) $$\mathfrak{F}^{(l)} \in \mathfrak{C}_b([0, +\infty[;\ \mathfrak{v}_{m-2l-1}),$$

then every solution of (1.1.7) *such that*

(2.8) $$\phi \in L^\infty(t_0, +\infty;\ \mathfrak{v}), \qquad t_0 \geqq 0,$$

satisfies

(2.9) $$\phi^{(j)} \in \mathfrak{C}_b([t_1, +\infty[;\ \mathfrak{E}_{m-2j}) \quad \forall t_1 > t_0, \quad j = 0, \cdots, l.$$

The notation $\mathfrak{C}_b([a,b[;\ X)$ means $\mathfrak{C}([a,b[;\ X) \cap L^\infty(a,b;\ X)$ and the proof of this theorem shows that the norm of the $\phi^{(j)}$ in $\mathfrak{C}_b([t_1, +\infty[,\ \mathfrak{E}_{m-2j})$ depends on the operators $\mathfrak{A}$, $\mathfrak{B}$, $\mathfrak{R}$, on the different norms of the $\mathfrak{F}^{(j)}$ in the spaces appearing in (1.2.6)–(1.2.7), on the norm of ϕ in $L^\infty(t_0, +\infty;\ \mathfrak{v})$ on j, t_0 and t_1.

Remark 2.1. When $\mathfrak{A} + \mathfrak{R}$ is $\mathfrak{v}$-coercive (i.e. there exists $\eta > 0$ such that $((\mathfrak{A} + \mathfrak{R})\phi, \phi) \geqq \eta \|\phi\|^2$, $\forall \phi \in \mathfrak{v}$) and when (1.1.13) is satisfied, it is shown in [16, p. 35] that the strong solution of (1.1.7)–(1.1.8) satisfies (1.2.8) with $t_0 = 0$.

1.3. Fractal dimension of an attractor. This section contains the main results of the first part. It is organized as follows. In a first time we show a squeezing property for the flow of strong solutions in the sense of C. Foias and R. Temam [14] (Theorem 1.3.1). Then we define functional invariant sets and attractors. We prove for these sets a regularity like property (Theorem 1.3.2) and finally we establish (Theorem 1.3.3) that invariant sets, bounded in $\mathfrak{v}$, have finite fractal dimension. In the previous works on the dimension of attractors ([14], [11], [4], [26], [23], $\cdots$) the authors have proved that the Hausdorff dimension of these sets is finite. P. Constantin, C. Foias and R. Temam [10] are the first who have shown that, for Navier–Stokes equations with homogeneous boundary conditions, the fractal dimension of attractors is finite (note that this notion of dimension is stronger than that of the Hausdorff dimension).

Squeezing property. We first introduce some notations. The operator $\mathfrak{A}$ is an isomorphism from $D(\mathfrak{A})$ onto $\mathfrak{H}$; $\mathfrak{A}^{-1}$ is a self-adjoint and compact operator on $\mathfrak{H}$. Thus $\mathfrak{A}$ possesses an orthonormal family of eigenvectors $\{\xi_j\}_{j \geqq 1}$, which is complete in $\mathfrak{H}$. The increasing sequence of eigenvalues of $\mathfrak{A}$ is denoted by $\{\lambda_j\}_{j \geqq 1}$ and we also denote by P_m the projector in $\mathfrak{H}$ onto the space spanned by $\xi_1, \cdots, \xi_m$. When $\dim \mathfrak{v} = \infty$ (we shall only consider this case) the sequence λ_j is infinite and

(3.1) $$\lim_{j \to +\infty} \lambda_j = +\infty.$$

In this section we assume that $\mathfrak{R}$ is linear and continuous from $\mathfrak{v}$ onto $\mathfrak{H}$ (i.e. $\theta_1=\theta_2=0$ in (1.1.3), (1.1.4)):

$$|\mathfrak{R}\psi|\leqq K_1\|\psi\| \quad \forall\psi\in\mathfrak{v}. \tag{3.2}$$

Let ϕ_1, ϕ_2 be two strong solutions of (1.1.7) on some interval $[0,T]$; $i=1,2$:

$$\frac{d\phi_i}{dt}+\mathfrak{A}\phi_i+\mathfrak{B}(\phi_i,\phi_i)+\mathfrak{R}\phi_i=\mathfrak{F}, \tag{3.3}_i$$

$$\phi_i(0)=\phi_i^0, \tag{3.4}_i$$

where ϕ_i^0 is given in $\mathfrak{v}$ and

$$\mathfrak{F}\in L^\infty(0,T;\ \mathfrak{H}). \tag{3.5}$$

Since strong solutions belong to $\mathfrak{C}([0,T];\ \mathfrak{v})$, we set

$$R=\max_{i=1,2}\Big(\sup_{t\in[0,T]}\|\phi_i(t)\|\Big). \tag{3.6}$$

THEOREM 3.1. *Under the previous hypotheses, there exist C_1 and C_2 depending only on $\mathfrak{A}$, $\mathfrak{B}$, $\mathfrak{R}$, R, T and $\mathfrak{F}$ such that for every $m\geqq 1$ and every $t\in[0,T]$ the following alternative holds*:

$$\textit{either (i)}\ |\phi_1(t)-\phi_2(t)|\leqq\sqrt{2}\,|P_m(\phi_1(t)-\phi_2(t))|,$$
$$\textit{or (ii)}\ |\phi_1(t)-\phi_2(t)|\leqq C_1\exp(-C_2\lambda_{m+1}\cdot t)|\phi_1^0-\phi_2^0|.$$

The proof given here is inspired from P. Constantin, C. Foias and R. Temam [10].

It should be noted that the constant $\sqrt{2}$ appearing in (i) could be any real number strictly larger than 1.

In the following the positive constants C_i depend only on $\mathfrak{A}$, $\mathfrak{B}$, $\mathfrak{R}$, R, T and on the norm of $\mathfrak{F}$ in $L^\infty(0,T;\ \mathfrak{H})$ (denoted $|\mathfrak{F}|_\infty$). Before proving Theorem 3.1 we establish:

LEMMA 3.1. *Under the assumption of Theorem* 1.3.1; *if $\phi_1^0\neq\phi_2^0$ then $\phi_1(t)\neq\phi_2(t)$ for every $t\in[0,T]$ and for $0\leqq t\leqq\tau\leqq T$, we have*

$$\frac{\|\phi_1(t)-\phi_2(t)\|^2}{|\phi_1(t)-\phi_2(t)|^2}\geqq\frac{\|\phi_1(\tau)-\phi_2(\tau)\|^2}{|\phi_1(\tau)-\phi_2(\tau)|}\exp\big(-C_3(\tau-t)^{1-\theta_4}\big). \tag{3.7}$$

Proof. We set $\psi=\phi_1-\phi_2$. From $(1.3.3)_{i=1,2}$ we deduce

$$\frac{d\psi}{dt}+\mathfrak{A}\psi+\mathfrak{B}(\phi_1,\psi)+\mathfrak{B}(\psi,\phi_2)+\mathfrak{R}\psi=0, \tag{3.8}$$

$$\psi(0)=\phi_1^0-\phi_2^0. \tag{3.9}$$

Denoting by $B(t)$ the time dependent operator

$$B(t)\xi=\mathfrak{B}(\phi_1(t),\xi)+\mathfrak{B}(\xi,\phi_2(t))+\mathfrak{R}\xi \quad \forall\in\mathfrak{v}; \tag{3.10}$$

from (1.1.6), (1.3.2) and from the fact that strong solutions belong to $L^2(0,T;\ D(\mathfrak{A}))$, we deduce that

$$B(t)\in L^2(0,T;\ \mathfrak{L}(\mathfrak{v},\mathfrak{H})). \tag{3.11}$$

Now according to C. Bardos and L. Tartar [6] the equation

$$\frac{d\psi}{dt}+\mathfrak{A}\psi+B(t)\psi=0,$$

with $B(t)$ satisfying (1.3.11) possesses the backward uniqueness property (note that (1.3.8) is ill-posed for $t<0$). Then if $\psi(0)\neq 0$, $\psi(t)\neq 0$, $0\leqq t\leqq T$. This proves the first part of the lemma.

We can set $\Lambda(t)=\|\psi(t)\|^2/|\psi(t)|^2$. According to (1.3.8) and to the identity $(\mathfrak{A}\psi-\Lambda\psi,\psi)=0$ we deduce that

$$\frac{1}{2}\frac{d\Lambda}{dt}+\frac{|\mathfrak{A}\psi-\Lambda\psi|^2}{|\psi|^2}=-\frac{(\mathfrak{R}\psi+\mathfrak{B}(\phi_1,\psi)+\mathfrak{B}(\psi,\phi_2),\ \mathfrak{A}\psi-\Lambda\psi)}{|\psi|^2}. \tag{3.12}$$

Now thanks to (1.1.3) and (1.1.6) we can majorize the r.h.s. of (1.3.12) by

$$\left[K_1+K_4\left(\|\phi_1\|^{1-\theta_4}|\mathfrak{A}\phi_1|^{\theta_4}+\|\phi_2\|^{1-\theta_4}|\mathfrak{A}\phi_2|^{\theta_4}\right)\right]\frac{\|\psi\|}{|\psi|}\cdot\frac{|\mathfrak{A}\psi-\Lambda\psi|}{|\psi|}.$$

We denote by σ the term between the brackets, and then

$$\frac{1}{2}\frac{d\Lambda}{dt}+\frac{|\mathfrak{A}\psi-\Lambda\psi|^2}{|\psi|^2}\leqq\sigma\frac{\|\psi\|}{|\psi|}\cdot\frac{|\mathfrak{A}\psi-\Lambda\psi|}{|\psi|}\leqq\frac{1}{2}\sigma^2\Lambda+\frac{1}{2}\frac{|\mathfrak{A}\psi-\Lambda\psi|^2}{|\psi|^2}.$$

Thus we have

$$\frac{d\Lambda}{dt}\leqq\sigma^2\Lambda \tag{3.13}$$

which yields by integration

$$\Lambda(t)\geqq\Lambda(\tau)\exp-\int_\tau^t\sigma^2(s)\,ds. \tag{3.14}$$

We estimate now this last integral. Thanks to Young's inequality

$$\int_\tau^t\|\phi_i\|^{2(1-\theta_4)}|\mathfrak{A}\phi_i|^{2\theta_4}ds\leqq\left(\int_\tau^t\|\phi_i\|^2ds\right)^{1-\theta_4}\left(\int_\tau^t|\mathfrak{A}\phi_i|^2ds\right)^{\theta_4},$$

$$\leqq(t-\tau)^{1-\theta_4}R^{2(1-\theta_4)}\left(\int_\tau^t|\mathfrak{A}\phi_i|^2ds\right)^{\theta_4}.$$

Thus in order to deduce (1.3.7) from (1.3.14) we must majorize $(\int_\tau^t|\mathfrak{A}\phi_i|^2ds)^{\theta_4}$. For that purpose we take the scalar product in $\mathfrak{H}$ of $(1.3.3)_i$ with $\mathfrak{A}\phi_i$:

$$\frac{1}{2}\frac{d}{dt}\|\phi_i\|^2+|\mathfrak{A}\phi_i|^2\leqq K_1\|\phi_i\|^{1-\theta_1}\cdot|\mathfrak{A}\phi_i|^{1+\theta_1}+K_4\|\phi_i\|^{2-\theta_4}\cdot|\mathfrak{A}\phi_i|^{1+\theta_4}+|\mathscr{F}|_\infty|\mathfrak{A}\phi_i|$$

(we have used (1.1.3) and (1.1.6)).

By applications of Young's inequality we deduce that

$$\frac{d}{dt}\|\phi_i\|^2+|\mathfrak{A}\phi_i|^2\leqq C_4,$$

from which we get by integration: $\int_0^T |\mathfrak{A}\phi_i|^2\,ds \leqq C_5$. Thus

$$\int_\tau^t \sigma^2(s)\,ds \leqq C_3(\tau - t)^{1-\theta_4}. \tag{3.15}$$

This completes the proof of Lemma 3.1.

Proof of Theorem 3.1. We take the scalar product in $\mathfrak{H}$ of (1.3.8) with ψ: (we use (1.1.5))

$$\frac{1}{2}\frac{d}{dt}|\psi|^2 + \|\psi\|^2 = -(\mathfrak{R}\psi,\psi) - ((\mathfrak{B}(\psi,\phi_2),\psi)$$

by (1.1.7) and (1.3.2)

$$\leqq \left[K_1 + K_4\|\phi_2\|^{1-\theta_4}|\mathfrak{A}\phi_2|^{\theta_4}\right]\|\psi\||\psi|$$

$$\leqq \frac{1}{2}\sigma^2|\psi|^2 + \frac{1}{2}\|\psi\|^2.$$

We obtain

$$\frac{d}{dt}|\psi|^2 + (\Lambda - \sigma^2)|\psi|^2 \leqq 0,$$

and thanks to (1.3.7) (τ is fixed)

$$\frac{d}{dt}|\psi|^2 + \left(\Lambda(\tau)\exp\left(-C_3\tau^{1-\theta_4}\right) - \sigma^2\right)|\psi|^2 \leqq 0.$$

By integration from 0 to τ we get

$$|\psi(\tau)|^2 \leqq |\psi(0)|^2 \exp\left(-\tau\Lambda(\tau)\exp\left(-C_3\tau^{1-\theta_4}\right)\right) + \int_0^\tau \sigma^2(s)\,ds\Big). \tag{3.16}$$

First case: $|Q_m\psi(\tau)| \leqq |P_m\psi(\tau)|$, then (i) is satisfied.
Second case: $|Q_m\psi(\tau)| > |P_m\psi(\tau)|$, then

$$\Lambda(\tau) \geqq \frac{\|P_m\psi(\tau)\|^2}{|Q_m\psi(\tau)|^2 + |P_m\psi(\tau)|^2} \geqq \frac{\lambda_{m+1}|Q_m\psi(\tau)|^2}{|Q_m\psi(\tau)|^2 + |Q_m\psi(\tau)|^2} = \frac{\lambda_{m+1}}{2}$$

and from (1.3.15) and (1.3.16) we deduce (ii) with τ instead of t. □

Remark 3.1. From (1.3.15), (1.3.16), and $\Lambda(\tau) \geqq \lambda_1$ we deduce the following property of continuous dependence w.r. to the initial value ϕ^0 for the flow $\{\phi(t)\}_{t\in[0,T]}$:

There exists a constant C_7 depending only on $\mathfrak{A}$, $\mathfrak{B}$, $\mathfrak{R}$, R, T and $|\mathscr{F}|_\infty$ such that

$$|\phi_1(t) - \phi_2(t)| \leqq C_7|\phi_1(0) - \phi_2(0)| \quad \forall t \in [0,T]. \tag{3.17}$$

Functional invariant sets, attractors. Let be given $\mathfrak{F} \in \mathfrak{H}$. For every $\phi^0 \in \mathfrak{v}$, as recalled in §1.1.1., the problem

Find $\phi \in \mathfrak{C}([0, T_1(\|\phi^0\|)];\ \mathfrak{v}) \cap L^2(0, T_1(\|\phi^0\|),\ D(\mathfrak{A}))$ such that

$$\frac{d\phi}{dt} + \mathfrak{A}\phi + \mathfrak{B}(\phi,\phi) + \mathfrak{R}\phi = \mathfrak{F}, \tag{3.18}$$

$$\phi(0) = \phi^0, \tag{3.19}$$

possess a unique solution which will be denoted by $S(t)\phi^0 \equiv \phi(t)$. This defines a nonlinear semi-group on $\mathfrak{v}$, continuous w.r. to the $\mathfrak{H}$ norm (see (1.3.17)).

DEFINITION 3.1. A functional invariant set X is a subset of $\mathfrak{v}$ such that

i) for every $\phi^0 \in X$, (1.3.18)–(1.3.19) possess a solution for all time (i.e. $\phi \in \mathfrak{C}([0, +\infty[; \mathfrak{v})$);

ii) $S(t)X = X$, $\forall t \geqq 0$.

Remark 3.2. As mentioned previously, the problem (1.3.18)–(1.3.19) is in general ill-posed for $t < 0$. Nevertheless for $\phi^0 \in X$, where X is a functional invariant set, this problem is well-posed for $t \in \mathbb{R}$. Indeed from (ii) of Definition 1.3.1 we deduce that for every $n \in \mathbb{N}$, there exist $\phi_n^0 \in X$ such that $S(n)\phi_n^0 = \phi^0$. Now for $t < 0$ and $n = [-t]$, we set $S(-t)\phi^0 = S(n+1-t)\phi_{n+1}^0$. Note that this definition is not ambiguous thanks to the backward uniqueness property pointed out at the beginning of the proof of Lemma 1.3.1. □

Examples.

(i) Let $\{\phi_1, \cdots, \phi_k\}$ be a set of steady-state solutions of (1.3.18); then $X = \{\phi_1, \cdots, \phi_k\}$ is functional invariant.

(ii) If $T > 0$ is such that $\phi(T) = \phi^0$ then the orbit $X = \{\phi(t),\ 0 \leqq t \leqq T\}$ is a functional invariant set.

(iii) Proposition 3.1 will exhibit a functional invariant set which is attracting in a certain sense. We do not know if this set is trivial (i.e. of the form (i)). □

An attractor is an invariant set attracting the trajectories which fall in its basin of attraction:

DEFINITION 3.2. An attractor is a functional invariant set X which possesses a neighbourhood in $\mathfrak{v}$, $\mathfrak{O}$, such that for every $\phi^0 \in \mathfrak{O}$, the distance from $S(t)\phi^0$ to X in $\mathfrak{v}$ tends to zero.

Remark 3.3. A. V. Babin and M. I. Vishik [4], [5] have proved the existence of an attractor in the case of two-dimensional Navier–Stokes equations in a rectangle with periodic boundary conditions (these equations can be written under the form (1.3.18)). In [4] it is showed that the dimension of that attractor grows at least like the inverse of the aspect ratio—the ratio of the length and the width of the strip.

Remark 3.4. The previous definitions were given in case of a constant excitation $\mathfrak{F}$. We can also consider the case of forced oscillation: $\mathfrak{F} \in L^\infty(0, +\infty; \mathfrak{H})$ and $\mathfrak{F}$ is T-periodic ($T > 0$). In this case we must substitute to the continuous semi-group $\{S(t)\}_{t \geqq 0}$ the distance one $\{S(nT)\}_{n \in \mathbb{N}}$ (see also Remark 3.7).

PROPOSITION 3.1. *Suppose that* $\mathfrak{B}$ *fulfills* (1.1.13) *and that* $\mathfrak{A} + \mathfrak{R}$ *is* $\mathfrak{v}$*-coercive* (*see Remark* 1.1.1). *For every* $\phi^0 \in \mathfrak{v}$ *and* $\mathfrak{F} \in \mathfrak{H}$, *there exists a functional invariant set* $X(\phi^0, \mathfrak{F})$, *bounded in* $D(\mathfrak{A})$, *such that*

$$\lim_{t \to +\infty} d_{\mathfrak{v}}\big(S(t)\phi^0, X\big) = 0. \tag{3.20}$$

Proof. As mentioned in Remark 1.1.1, under the above hypotheses, the problem (1.3.18)–(1.3.19) possesses a solution $\phi \in \mathfrak{C}([0, +\infty[; \mathfrak{v})$ and moreover $\phi \in L^\infty(0, +\infty; \mathfrak{v})$. We set $X(\phi^0, \mathfrak{F}) = \bigcap_{\tau \geqq 0} \{\overline{\phi(s),\ s > \tau}\}^{\mathfrak{H}}$, and remark that $\phi \in X$ is equivalent to the existence of a sequence of real numbers $(s_j)_{j \geqq 1}$ such that

$$\lim_{j \to +\infty} s_j = +\infty, \qquad \lim_{j \to +\infty} \big|S(s_j)\phi^0 - \psi\big| = 0. \tag{3.21}$$

Since $\phi \in L^\infty(0, +\infty; \mathfrak{v})$, we deduce from Theorem 1.2.1 with $m = 2$ that $\phi \in L^\infty(1, +\infty, \mathfrak{E}_2)$ and by (1.2.2) it results that X is bounded in $D(\mathfrak{A})$ (see also Remark 1.3.5).

We prove now (ii) of Definition 1.3.1. Let be $t>0$ and $\phi\in S(t)X$. There exist $\psi\in S$ and $(s_j)_{j\geq 1}$ satisfying $\phi=S(t)\psi$ and $\phi(s_j)\to\psi$ in $\mathfrak{H}$. By continuity of $S(t)$ we have $\phi(s_j+t)\to\phi$ in $\mathfrak{H}$ and thus $\phi\in X: S(t)X\subset X$. Conversely, let $\phi\in X$. There exists $(s_j)_{j\geq 1}$ such that $\phi(s_j)\to\phi$ in $\mathfrak{H}$. We set $\sigma_j=s_j-t$ where $t>0$ is fixed; the sequence $(\phi(\sigma_j))$ is bounded in $D(\mathfrak{A})$, by compactness we can extract from $(\phi(\sigma_j))$ a subsequence $(\phi(\sigma_{j'}))$ which converges to some limit ψ in $\mathfrak{H}$. We have $\phi(s_{j'})=S(t)\phi(\sigma_{j'})\to S(t)\psi$ in $\mathfrak{H}$ and since the sequence $(\phi(s_{j'}))$ is extracted from $(\phi(s_j))$ we have $\phi(s_{j'})\to\phi$ in $\mathfrak{H}$. Thus $S(t)\psi=\phi$ and then $\phi\in S(t)X: X\subset S(t)X$.

It remains to show that the distance in $\mathfrak{v}$ from X to $S(t)\phi^0$ tends to zero. If it were not the case, there would exist some $\varepsilon_0>0$ and a sequence (t_j), $t_j\to+\infty$, such that $d_{\mathfrak{v}}(\phi(t_j), X)\geq\varepsilon_0$. But the sequence $(\phi(t_j))$ is bounded in $D(\mathfrak{A})$ and thus relatively compact in $\mathfrak{v}$. We can extract a subsequence $(\phi(t_{j'}))$ which converges to some limit ψ w.r. to the $\mathfrak{v}$-norm and since (1.3.21) is satisfied, $\psi\in X$. The function $d_{\mathfrak{v}}(\cdot, X)$ is continuous on $\mathfrak{v}$; thus $d(\psi, X)\geq\varepsilon_0$ which contradicts $\psi\in X$.

Remark 3.5. Assume that $\mathfrak{F}\in\mathfrak{E}_{m_0}$ for some $m_0\geq 0$, and then according to Theorem 1.2.1 with $m=m_0$, $X(\phi^0,\mathfrak{F})$ is bounded in $\mathfrak{E}_{m_0+2}$ and

$$\lim_{t\to+\infty} d_{\mathfrak{E}_{m_0+1}}\big(S(t)\phi^0, X\big)=0.$$

Regularity. In all the applications that we have in view, the operator $\mathfrak{A}$ is in a certain sense an elliptic differential operator of even order. Thus the spaces $\mathfrak{v}_m$ (and $\mathfrak{E}_m$) are spaces of functions more regular as m increases (think that $\mathfrak{E}_m$ is approximately a space of functions whose mth distributional derivative is still a function). The next result shows that functional invariant sets (and thus attractors) are as regular as $\mathfrak{F}$.

THEOREM 3.2. *Let $\mathfrak{F}\in\bigcap_{m\geq 0}\mathfrak{E}_m$, and then every functional invariant set, X, is included in $\bigcap_{m\geq 0}\mathfrak{E}_m$. Moreover if X is bounded in $\mathfrak{v}$, then X is bounded in $\mathfrak{E}_m$ for every $m\geq 0$.*

Proof. Let $\psi\in X$, and then there exists $T>0$ and $\phi^0\in\mathfrak{v}$ such that ϕ, the solution of (1.3.18)–(1.3.19), satisfies $\phi(T)=\psi$ and $\phi\in C([0,T];\ \mathfrak{v})$. Then according to R. Temam [30], we have $\phi(T)\in\mathfrak{E}_m$, for every $m\in\mathbb{N}$.

If X is bounded in $\mathfrak{v}$, then for every $\phi^0\in X$ the flow $\phi(t)$ satisfies (1.2.8) with $t_0=0$. Thus Theorem 1.2.1 applies for every $m\in\mathbb{N}$ and ϕ is bounded in $\mathfrak{E}_m$ by a constant which does not depend on ϕ^0. This proves the last part of Theorem 3.2. □

Remark 3.6. Assuming only that $\mathfrak{F}\in\mathfrak{E}_{m_0}$, for some $m_0\in\mathbb{N}$, we obtain that $X\subset\mathfrak{E}_{m_0+2}$. And when X is bounded in $\mathfrak{v}$, X is bounded in $\mathfrak{E}_{m_0+2}$.

Finite dimension. We show that functional invariant sets, bounded in $\mathfrak{v}$, have finite dimension. We first introduce the definition of fractal dimension (see B. Mandelbrot [25]):

DEFINITION 3.3. Let Y be a subset of a metric space X. For every $\delta>0$, the minimal number of open balls of radius equals to δ which is necessary to cover Y is denoted by $N_Y(\delta)$ (this number can be infinite). We set ($\gamma\in\mathbb{R}_+^*$):

$$\mu_F(Y,\gamma)=\limsup_{\varepsilon\to 0} N_Y(\varepsilon)\varepsilon^\gamma.$$

If there exists $\delta\geq 0$ such that $\mu_F(Y,\gamma)<+\infty$, Y has finite fractal dimension and its dimension is the number

$$\inf\{\gamma\geq 0,\ \mu_F(Y,\gamma)<+\infty\}.$$ □

THEOREM 3.3. *Every functional invariant set, bounded in* $\mathfrak{v}$, *has finite fractal dimension.*

Proof. Let be R such that $\|\phi\| \leqq R$ for every ϕ in X, the functional invariant set considered here. We also fix $T>0$.

According to Theorem 1.3.1, for every ϕ and ψ in $\mathfrak{v}$ we have

$$(3.22)\quad |S(T)\phi - S(T)\psi| \leqq C_1 \exp(-C_2\lambda_{m+1}T)|\phi-\psi| + \sqrt{2}\,|P_m(S(T)\phi - S(T)\psi)| \quad \forall \phi, \psi \in \mathfrak{v}, \|\phi\| \leqq R, \|\psi\| \leqq R.$$

And thanks to (1.3.17)

$$(3.23)\quad |S(T)\phi - S(T)\psi| \leqq C_7|\phi-\psi| \quad \forall \phi, \psi \in \mathfrak{v}, \|\phi\| \leqq R, \|\psi\| \leqq R.$$

Let $\eta \in]0,1[$, and from (1.3.1) and (1.3.22) we deduce that there exist a constant C_8 $(=\sqrt{2})$ and a continuous projector on $\mathfrak{H}$ of finite dimension Π $(=P_m$, with m such that $\eta \leqq C_1 \exp(-C_2\lambda_{m+1}t))$ such that

$$(3.24)\quad |S(T)\phi - S(T)\psi| \leqq \eta|\phi-\psi| + C_8|\Pi(S(T)\phi - S(T)\psi)|.$$

We shall prove that Theorem 1.3.3 is a consequence of (1.3.23) and (1.3.24).

Let there be given $\gamma>0$ and $\varepsilon>0$. Since X is relatively compact in $\mathfrak{H}$, it can be covered by a finite number of open balls $\{B_i\}_{1 \leqq i \leqq N}$ of radius ε. The ball B_i is supposed centered at $\phi_i \in X$. From $S(T)X \supset X$ we deduce that $X \subset \bigcup_{i=1}^{N} S(T)(B_i \cap X)$.

Let there be given ϕ in $S(T)(B_i \cap X)$, and thanks to (1.3.23) and (1.3.24) we have $(\psi_i = S(T)\phi_i)$

$$(3.25)\quad |\phi - \psi_i| \leqq \eta\varepsilon + C_8|\Pi(\phi - \psi_i)|,$$

$$(3.26)\quad |\phi - \psi| \leqq C_7\varepsilon.$$

We construct a new covering of X. For i fixed, according to (1.3.26), $\Pi(S(T)(B_i \cap X))$ is included in a ball of $\Pi(\mathfrak{H})$ of radius $2C_7\varepsilon$. Let be $\tilde{B}_i^1, \cdots, \tilde{B}_i^P$ a minimal covering of $\Pi(S(t)(B_i \cap X))$ by balls of radius $r_1 = (1-\eta)/4C_8\varepsilon$. We have $p \leqq l_m(r_1/2C_7\varepsilon) = l_m((1-\eta)/2C_7C_8)$, where $l_m(\rho)$ is the minimal number of balls of radius ρ which is necessary to cover the unit ball of $\mathbb{R}^m$. We set $G_i^k = S(T)(B_i \cap X) \cap \Pi^{-1}(\tilde{B}_i^k)$. According to (1.3.25), G_i^k is included in a closed ball, centered at ψ_i, of radius $\leqq \eta\varepsilon + C_8 r_1 = (1+3\eta)/4\varepsilon$. Thus we can include G_i^k in an open ball B_i^k of radius $(1+\eta)\varepsilon/2$. Let us now summarize what we have done. Starting from a generic covering of X by balls of radius ε, $\{B_i\}_{1 \leqq i \leqq N}$, we have constructed a new covering $\{B_i^k\}_{1 \leqq i \leqq N, 1 \leqq k \leqq \rho}$, by balls of radius $(1+\eta)\varepsilon/2$ $(<\varepsilon)$. Thus we have: $N_X((1+\eta)\varepsilon/2) \leqq p \cdot N_X(\varepsilon)$. Denoting by $\phi(\varepsilon) = N_X(\varepsilon)\varepsilon^\gamma$, we get from this last inequality that $\phi((1+\eta)/2\varepsilon) \leqq l_m(\eta/8C_7C_8) \cdot ((1+\eta)/2)^\gamma \phi(\varepsilon)$. We claim that if $l_m((1-\eta)/8C_7C_8)((1+\eta)/2)^\gamma < 1$ i.e. if

$$(3.27)\quad \gamma > \gamma_0 = \frac{\log l_m((1-\eta)/8C_4C_5)}{\log 2/(1+\eta)},$$

we have $\limsup_{\varepsilon \to 0} \phi(\varepsilon) = 0$ and thus the fractal dimension of X is not greater than γ_0.

It remains to prove our claim. We set $\delta = (1+\eta)/2$ and $\chi = l_m((1-\eta)/8C_4C_5) \cdot ((1+\eta)/2)^\gamma$. We have $\phi(\delta\varepsilon) \leqq \chi\phi(\varepsilon)$ thus by iteration $\phi(\varepsilon) \leqq \chi^j \phi(\delta^{-j}\varepsilon)$, $\forall j \in \mathbb{N}$.

Taking $j = 1 + [\log\varepsilon/\log\delta]$, $\phi(\delta^{-j}\varepsilon) \leqq \sup_{1 \leqq \rho \leqq \delta^{-1}} \phi(\rho) \leqq \delta^{-\gamma} N_X(1) < +\infty$ and $\phi(\varepsilon) \leqq \chi^j \delta^{-\gamma} N_X(1)$ tends to 0 with ε. □

Remark 3.7. (Continuation of Remark 1.3.4): Theorems 1.3.1, 1.3.2 and 1.3.3 are also valid in case of forced oscillations.

2. Applications. In this part we illustrate the abstract results of the first part by four applications in fluid mechanics. The four following sections are set in the same way: after writing the set of equations coming from physics, we give the functional setting that allows us to use the results of the first part. Then we state in each particular situation the results obtained in §1.3.

2.1. Nonhomogeneous Navier–Stokes equations. Many of the results obtained here were proved by C. Foias and R. Temam [14]. Nevertheless we need to present them in order to introduce some notations and to check the hypotheses of the abstract framework, which will be subsequently useful.

We consider the motion of a viscous incompressible fluid which fills some bounded region Ω. The velocity $u(x,t)$ and pressure $p(x,t)$, are determined by the set of equations

$$\begin{aligned} &\frac{\partial u}{\partial t}-\nu\Delta u+(u\cdot\nabla)u+\nabla p=h && \text{in } \Omega\times\mathbb{R}_+,\\ &\operatorname{div} u=0 && \text{in } \Omega\times R_+,\\ &u(x,t)=\psi(x) && \text{on } \Gamma\times\mathbb{R}_+,\\ &u(x,0)=u^0(x) && \text{in } \Omega; \end{aligned} \tag{1.1}$$

where ν is the viscosity of the fluid, h is an external force acting on the fluid and $\psi(x)$ is the velocity of the boundary Γ of Ω (we have taken the density equal to 1).

2.1.1. Functional setting. We assume that Ω is an open connected bounded set in $\mathbb{R}^d$ ($d=2$ or 3) and that

$$\Gamma \text{ is a } C^\infty\text{-manifold of dimensional } d-1 \text{ and } \Omega \text{ is locally located on one side of } \Gamma, \tag{1.2}$$

$$\Gamma \text{ has a finite number of connected components denoted } \Gamma_1,\cdots,\Gamma_k. \tag{1.3}$$

Let H and V be the closures in $L^2(\Omega)^d$ and $H_0^1(\Omega)^d$ respectively of

$$\mathcal{N}=\left\{v\in\mathfrak{D}(\Omega)^d,\ \operatorname{div} v=0\right\}$$

where $H_0^m(\Omega)$ denotes the closure of $\mathfrak{D}(\Omega)$ (the space of C^∞ functions with compact support in Ω) in the Sobolev L^2-space $H^m(\Omega)$ of order m.

The spaces H and V are endowed with the scalar products

$$(u,v)=\int_\Omega u_i v_i dx \quad\text{and}\quad ((u,v))=\int_\Omega D_j u_i D_j v_i dx$$

respectively. We set also $|u|=(u,u)^{1/2}$ and $\|u\|=((u,u))^{1/2}$. The injection from V into H is dense, continuous and compact (thanks to Rellich's lemma). Let A be the operator from V into V' defined by

$$\langle Au,v\rangle=((u,v)).$$

We set

$$E_0 = H \quad \text{and} \quad E_m = H^m(\Omega)^d \cap H, \qquad m \geqq 1 \tag{1.4}$$

and according to the regularity properties of Stokes operator A (see R. Temam [32], J. M. Ghidaglia [17]):

$$A^{-1} \text{ is a continuous operator from } \mathfrak{E}_m \text{ into } \mathfrak{E}_{m+2} \cap V. \tag{1.5}$$

For u, v, w in $H^1(\Omega)^d$ we set

$$b(u,v,w) = \int_\Omega u_j (D_j v_k) w_k \, dx. \tag{1.6}$$

We have the following property for b (see R. Temam [31]):

$$\begin{aligned} &b \text{ is continuous on } H^{s_1}(\Omega)^d \times H^{s_2+1}(\Omega)^d \times H^{s_3}(\Omega)^d, \text{ where } s_i \geqq 0 \text{ and} \\ &s_1 + s_2 + s_3 > \frac{d}{2} \quad \text{if one of the } s_i = \frac{d}{2}, \\ &s_1 + s_2 + s_3 \geqq \frac{d}{2} \quad \text{otherwise.} \end{aligned} \tag{1.7}$$

Let $u, v, w \in V$, we set $B(u,v) \in V' : \langle B(u,v), w \rangle = b(u,v,w)$, then

$$B \text{ is continuous from } E_{m+1} \times E_{m+1} \text{ into } E_m \text{ for } m \geqq 1. \tag{1.8}$$

The following result (see R. Temam [32] for the proof) will be useful to find an appropriate $\bar{u} \in H^1(\Omega)^d$ such that $\bar{u}_{|\Gamma} = \psi$.

LEMMA 1.1. *Let $\psi \in H^{1/2}(\Gamma)^d$ be given such that (n is the unit outward normal on Γ)*

$$\int_{\Gamma_i} \psi \cdot n \, d\Gamma = 0, \qquad i = 1, \cdots, k. \tag{1.9}$$

There exists $\bar{u} \in H^1(\Omega)$ such that

i) $\bar{u}_{|\Gamma} = \psi$ *(in the sense of traces on Γ),*

ii) $\operatorname{div} \bar{u} = 0$,

iii) $|b(v, \bar{u}, v)| \leqq \nu \|v\|^2/2$, $\forall v \in V$.

Moreover, for every $m \in \mathbb{N}$, if $\psi \in H^{m+1/2}(\Gamma)^d$, $\bar{u} \in H^{m+1}(\Omega)^d$.

Using now the classical projection on divergence free fields (see J. L. Lions [24], R. Temam [32], $\cdots$) we deduce that (2.1.1) is equivalent to the initial value problem for $\phi = u - \bar{u}$ ($\bar{u}$ is given by Lemma 2.1.1):

$$\frac{d\phi}{dt} + \nu A\phi + B(\phi,\phi) + R\phi = f, \tag{1.10}$$

$$\phi(0) = \phi^0, \tag{1.11}$$

where $R \in \mathfrak{L}(V, V')$ is defined by

$$\langle R\phi, \xi \rangle = b(\phi, \bar{u}, \xi) + b(\bar{u}, \phi, \xi) \quad \forall \xi \in V. \tag{1.12}$$

2.1.2. Fractal dimension for attractors. We briefly show that the operators introduced previously satisfy the assumptions of the first part.

Suppose that $\psi \in H^{3/2}(\Gamma)$ satisfies (2.1.9); then according to Lemma 2.1.1, (2.1.7) and (2.1.12), R satisfies (1.3.2) and B satisfies (1.1.6) with $\theta_3 = \frac{1}{2}$. Note that (1.1.5) is the well-known property of orthogonality of b. The estimate (1.1.7), with $\theta_4 = \frac{1}{2}$, results from (2.1.7) and from (cf. S. Agmon [1]):

$$|\phi|^2_{L^\infty(\Omega)} \leqq C\|\phi\||A\phi| \quad \forall \phi \in D(A).$$

Remark 1.1. When $\Omega \subset \mathbb{R}^2$, according to (2.1.7), (1.1.13) is satisfied with $\theta_5 = \frac{1}{2}$. On the other hand, thanks to the choice of $\bar{u}$ (Lemma 2.1.1):

$$((\nu A + R)\phi, \phi) \geqq \frac{\nu}{2}\|\phi\|^2 \quad \forall \phi \in V.$$

As mentioned in Remark 1.2.1, the problem (2.1.10)–(2.1.11) possesses a unique strong solution for $t \in [0, +\infty[$ which is bounded in $H^1(\Omega)^2$. Note also that, according to Proposition 1.3.1, for every $u^0 \in V$ and $h \in L^2(\Omega)^d$, there exists a functional invariant set $X(u^0, h)$, bounded in $H^2(\Omega)^d$, such that the distance from $u(t)$ to X in $H^1(\Omega)^d$ tends to zero when t goes to $+\infty$. □

THEOREM 1.1. *Assume that $h \in \mathfrak{C}^\infty(\bar{\Omega})^d$ and $\psi \in \mathfrak{C}^\infty(\Gamma)^d$; then every functional invariant set, X, lies in $\mathfrak{C}^\infty(\bar{\Omega})^d$. Moreover if X is bounded in $H^1(\bar{\Omega})^d$, then X is bounded in $H^m(\Omega)^d$ for every $m \geqq 0$.*

Proof. From the hypotheses on h and ψ we deduce that $f \in \bigcap_{m \geqq 0} E_m$ and $\bar{u} \in \mathfrak{C}^\infty(\bar{\Omega})^d$. It follows then that R maps E_{m+1} into E_m; thus Theorem 1.3.2 applies.

From Theorem 1.3.3 we deduce:

THEOREM 1.2. *Every functional invariant set, bounded in $H^1(\Omega)^d$, has finite fractal (and Hausdorff) dimension.*[1]

2.2. Navier–Stokes equations on a manifold. This application is motivated by the modelisation of geophysical flows. Assuming that the velocities are horizontal and proportional to the distance from the center of the Earth, this flow can be viewed as a viscous incompressible flow on the 2D-sphere S^2 (see for instance A. Avez and Y. Bamberger [3]). In the following analysis this case corresponds to $M = S^2$ and the metric g is that induced by the 3-Euclidean space. Note that in this case the following assumption (2.2.4) is a well-known property of sphere with even dimension.

It should be observed that Navier–Stokes equations with periodic boundary conditions correspond to $M = T^n$ (n-dimensional torus) endowed with the flat metric ($g_{ij} = \delta_{ij}$).

2.2.1. Geometric preliminaries. Let M be an n-dimensional compact, connected and oriented Riemannian manifold without boundary.

We denote by g the Riemannian metric on M: g is a C^∞-field of symmetric and bilinear forms on the tangent spaces:

$$g(x; u, v) = g(x; v, u) \quad \forall x \in M, \quad \forall u, \quad v \in T_x M$$

where $T_x M$ is the tangent space at M in x,

$$g(x; u, u) > 0 \quad \forall x \in M, \quad \forall u \in T_X M, \quad u \neq 0.$$

[1] Recall that if a set has finite fractal dimension then it has finite Hausdorff dimension. The converse is false.

Letting $(x^1, \cdots, x^n)$ be a coordinate system on M, we set

$$g_{ij} = g\left(x; \frac{\partial}{\partial x^i}, \frac{\partial}{\partial x^j}\right), \qquad G = \det(g_{ij}),$$

and denote by

$$(u \mid v) = g_{ih} u^i u^j$$

the scalar product on $T_X M$ induced by g and $|\cdot|$ the corresponding norm.

The inverse of g_{ij} is denoted by $g^{ij}: g_{ij} g^{jk} = \delta_i^k$. Thanks to g_{ij} and g^{ij} we can:
assign to u (vector), $\tilde{u}$ (covector) by $(\tilde{u})_i = g_{ij} u^j$,
assign to ω (covector), $\hat{\omega}$ (vector) by $(\hat{\omega})^i = g^{ij} \omega_j$.

In the sequel we shall omit the $\sim$ and the $\wedge$ and use subscripts for covariant index and superscripts for contravariant index.

2.2.2. Differential operators, Sobolev spaces. We denote by $C^\infty(M)$ (resp. $C^\infty(TM)$) the space of C^∞-functions (resp. C^∞-vector field) on M.

Gradient of a function. We associate to $p \in C^\infty(M)$, the element $\nabla p \in C^\infty(TM)$:

$$(\nabla p)^i = g^{ij} \frac{\partial p}{\partial x^j}.$$

Total derivative of a vector field. Let $u \in C^\infty(TM)$ and D the Levi–Cività connection (see (2.2.1) for the expression in a coordinate system),

$$(\nabla u)_i^j = D_i u^j.$$

Divergence of a vector field. The divergence is the contraction of the total derivative

$$\operatorname{div} u = D_i u^i.$$

Laplacian on $\mathfrak{C}(TM)$.

$$(\Delta u)^i = g^{kl} D_k D_l u^i \quad \forall u \in C^\infty(TM).$$

In order to give the expression of these differential operators in a coordinate system, we introduce Christoffel's symbols:

$$\{i \quad j \quad k\} = \frac{1}{2}\left(\frac{\partial g_{jk}}{\partial x^i} + \frac{\partial g_{ik}}{\partial x^j} - \frac{\partial g_{ij}}{\partial x^k}\right),$$

$$\left\{\begin{matrix} k \\ i\, j \end{matrix}\right\} = g^{kl}\{ijl\}.$$

The expression of the covariant derivation is then

$$D_i u^j = \frac{\partial u_j}{\partial x^i} + \left\{\begin{matrix} j \\ i\, k \end{matrix}\right\} u^k. \tag{2.1}$$

2.2.3. Navier–Stokes equations on M. With the previous notation, Navier–Stokes equations on M read (see D. Ebin and J. Marsden [12]):

$$\frac{\partial u}{\partial t} - \nu \Delta u + D_u u + \nabla p = f, \tag{2.2}$$

$$\operatorname{div} u = 0, \tag{2.3}$$

where f is a tangent vector field, the unknowns u (the velocity) and p (the pressure) being respectively a tangent vector field and a scalar field.

Let us introduce Stokes operator on M. We set

$$\mathcal{N}=\{u\in C^\infty(TM),\ \operatorname{div}u=0\},$$

and let H and V be the closures in $\mathbb{L}^2$ and $\mathbb{H}^1$ respectively of $\mathfrak{v}$ (for Sobolev spaces on manifolds the reader is referred to T. Aubin [2]). These spaces are endowed with the scalar products

$$(u,v)=\int(uv)\,dM \qquad (dM=\sqrt{G}\,dx^1\cdots dx^n)$$

and

$$((u,v))_1=\int(u\,|\,v)\,dM+\int(\nabla u\,|\,\nabla v)\,dM$$

respectively. We set also $\|u\|_1=((u,u))_1^{1/2}$.

Let A be the operator from V into V' defined by

$$\langle Au,v\rangle=\int(\nabla u\,|\,\nabla v)\,dM \quad \forall u,\ v\in V.$$

The next proposition gives a condition which is sufficient to insure the coercivity of A:

PROPOSITION 2.1. *The semi-norm*

$$|u|_1=\left(\int|\nabla u|^2\,dM\right)^{1/2}$$

reduces to a norm equivalent to $\|\cdot\|_1$ *on* $\mathbb{H}^1$ *if the metric* g *does not locally split under the form* $g'+dy_n^2$ *where* g' *depends only on* $n-1$ *variables.* □

This proposition is a consequence of the de Rham decomposition theorem (see S. Kobayashi and K. Nomizu [21]). For further details see [16]. It is also proved in this last reference that under the topological condition:

(2.4) Every continuous vector field on M, vanishes at least at one point,

the semi-norm $|\cdot|_1$ is equivalent to $\|\cdot\|_1$.

Let v be given in $\mathfrak{v}$. The scalar product in $\mathbb{L}^2$ of (2.2.2) with v yields since $\operatorname{div}v=0$:

$$\frac{\partial}{\partial t}(u\,|\,v)+\nu\langle Au,v\rangle+\langle B(u,u),\,v\rangle=(f\,|\,v),$$

where we set

$$\langle B(u,v),\,w\rangle=\int(D_u v\,|\,w)\,dM \quad \forall u,\ v,\ w\in\mathfrak{E}(TM).$$

With this notation, (2.2.2)–(2.2.3) turn out to be equivalent to the differential equation for u:

$$\frac{du}{dt}+\nu Au+B(u,u)=f. \tag{2.5}$$

2.2.4. Fractal dimension of attractors. The verification of the fact that the previous functional setting is compatible with the hypotheses of the first part (when dimension of $M=2$ or 3) is similar to that of the case of Navier–Stokes equations (here $\mathfrak{R}\equiv 0$). The details are given in [16].

Remark 2.1. When $\dim M=2$, (1.1.13) is satisfied with $\theta_5=\frac{1}{2}$. According to Remark 1.2.1, for every $f\in L^\infty(0,+\infty;\ H)$ and $u^0\in V$ the initial value Problem for (2.2.5) with $u(0)=u^0$ possesses a strong solution which is bounded in $\mathbb{H}^1$. Note also that Proposition 1.3.1 applies in this case. □

We deduce from the results of §1.3:

THEOREM 2.1. *Assume that $f\in C^\infty(TM)$; then every functional invariant set, X, lies in $C^\infty(TM)$. Moreover, if X is bounded in $\mathbb{H}^1$, then X is bounded in $\mathbb{H}^m$ for every $m\geqq 0$.*

THEOREM 2.2. *Every functional invariant set, bounded in $\mathbb{H}^1$, has finite fractal (and Hausdorff) dimension.*

2.3. Thermo-hydraulic equations. We consider the motion of a viscous incompressible fluid, subjected to thermal effects, which fills some bounded region Ω. In the Boussinesq approximation the velocity $u(x,t)$, pressure $p(x,t)$ and temperature $\theta(x,t)$ are determined, in case of homogeneous boundary conditions, by the equations (see Chandrasekhar [7]):

$$
\begin{aligned}
&\frac{\partial u}{\partial t}+(u\cdot\nabla)u-\nu\Delta u+\sigma\theta+\nabla p=f,\\
&\frac{\partial \theta}{\partial t}+(u\cdot\nabla)\theta-\kappa\Delta\theta=s,\\
&\operatorname{div} u=0,\\
&u(x,t)=0 \quad \text{on } \partial\Omega\times\mathbb{R}_+,\\
&\theta(x,t)=0 \quad \text{on } \partial\Omega\times\mathbb{R}_+;
\end{aligned}
\tag{3.1}
$$

where κ denotes the Fourier coefficient of the fluid and σ is a fixed vector in $\mathbb{R}^d$ pointing vertically downwards.

2.3.1. Functional setting. We supplement some notations to that of the §2.1.1. The orthogonal projector in $L^2(\Omega)^d$ onto H is denoted by P. The operator $-\Delta$ on $H_0^1(\Omega)$ is denoted by A_1 and we define $B_1\in\mathfrak{L}(V\times H_0^1(\Omega),\ H^{-1}(\Omega))$ by

$$\langle B_1(U,\theta),\ \eta\rangle=\int_\Omega u_i\frac{\partial\theta}{\partial x_i}\eta\,dx.$$

We set

$$\mathfrak{H}=H\times L^2(\Omega),\qquad \mathfrak{v}=V\times H_0^1(\Omega);$$

these spaces are endowed with the scalar products

$$(\phi_1,\phi_2)=\int_\Omega(u_1\cdot u_2+\theta_1\cdot\theta_2)\,dx,$$

and

$$((\phi_1,\phi_2))=\nu\int_\Omega\nabla u_1:\nabla u_2\,dx+\kappa\int_\Omega\nabla\theta_1\cdot\nabla\theta_2\,dx$$

respectively, where ϕ_i denotes the pair $\{u_i,\theta_i\}$.

We introduce the operators $\mathfrak{A}\in\mathfrak{L}(\mathfrak{v},\mathfrak{v}')$, $B\in\mathfrak{L}(\mathfrak{v}\times\mathfrak{v},\mathfrak{v}')$ and $\mathfrak{R}\in\mathfrak{L}(\mathfrak{v},\mathfrak{v}')$:

$$\langle \mathfrak{A}\phi_1,\phi_2\rangle=((\phi_1,\phi_2)),$$

$$\langle \mathfrak{B}(\phi_1,\phi_2),\phi_3\rangle=\int_\Omega (u_1\cdot\nabla)u_2u_3\,dx+\int_\Omega (u_1\cdot\nabla)\theta_2\theta_3\,dx,$$

$$\langle \mathfrak{R}\phi_1,\phi_2\rangle=\int_\Omega \theta_1\sigma\cdot u_2\,dx.$$

With these notations and $\mathfrak{F}=(f,s)$ equations (2.3.1) read

$$\frac{d\phi}{dt}+\mathfrak{A}\phi+\mathfrak{B}(\phi,\phi)+\mathfrak{R}\phi=\mathfrak{F}.$$

2.3.2. Fractal dimension of attractors. The verification of the fact that the previous functional setting is compatible with the hypotheses of the first part is similar to that of the case of Navier–Stokes equation. The details are given in [16].

Remark 3.1. When $\Omega\subset\mathbb{R}^2$, (1.1.3) is satisfied with $\theta_5=\frac{1}{2}$. In general $\mathfrak{A}+\mathfrak{R}$ is not $\mathfrak{v}$-coercive, but studying the system equivalent to (2.3.1), obtained by multiplying the evolution equation of θ by a large constant, we change $\mathfrak{A}$ to make $\mathfrak{A}+\mathfrak{R}$ $\mathfrak{v}$-coercive. Thus Proposition 1.3.1 applies also in this case. Note that in case of nonhomogeneous boundary condition on the temperature, the trick mentioned previously does not apply but, as proved in [16], the conclusions of Remark 1.2.1 and Proposition 1.3.1 still hold. □

We deduce from the results of §1.3:

THEOREM 3.1. *Assume that* $f\in\mathfrak{C}^\infty(\bar\Omega)^d$ *and* $s\in\mathfrak{C}^\infty(\bar\Omega)$; *then every functional invariant set,* X, *lies in* $\mathfrak{C}^\infty(\bar\Omega)^{d+1}$. *Moreover if* X *is bounded in* $H^1(\Omega)^{d+1}$, *then* X *is bounded in* $H^m(\Omega)^{d+1}$ *for every* $m\geqq 0$.

THEOREM 3.2. *Every functional invariant set, bounded in* $H^1(\Omega)^d$, *has finite fractal (and Hausdorff) dimension.*

2.4. Magnetohydrodynamic equations. In this section we briefly give the functional setting that allows us to apply the results of the first part to M. H. D. equations. Most of the results which issue from this application are contained in M. Sermange and R. Temam [29]. Nevertheless we have presented them in order to emphasize the generality of the abstract framework.

We consider the motion of a viscous incompressible and resistive fluid. The velocity $u(x,t)$, pressure $p(u,t)$ and the magnetic field $B(x,t)$ are determined by the equations (see L. Landau and E. Lifchitz [22]):

$$\begin{aligned}
&\frac{\partial u}{\partial t}+(u\cdot\nabla)u-\frac{1}{Re}\Delta u-(B\cdot\nabla)B+\nabla\left(p+\frac{1}{2}B^2\right)=f,\\
&\frac{\partial B}{\partial t}+(u\cdot\nabla)B+\frac{1}{Rm}\operatorname{curl}(\operatorname{curl}B)-(B\cdot\nabla)u=0,\\
&\operatorname{div}u=0,\qquad \operatorname{div}B=0,\\
&u=0,\quad B\cdot n=0\ \text{ and }\ (\operatorname{curl}B)\times n=0\quad\text{on }\Gamma,
\end{aligned}\tag{4.1}$$

where n is the unit normal on Γ. The two positive numbers Re and Rm appearing in (2.4.1) are respectively the Reynolds number and the magnetic Reynolds number.

We assume, in addition to the hypothesis of §2.1.1 that Ω is simply connected (this assumption is not essential) and introduce the space

$$\tilde V=\left\{B\in H^1(\Omega)^d,\ \operatorname{div}B=0,\ B\cdot n=0\right\}.$$

We set $\mathfrak{H}=H\times H$ and $\mathfrak{v}=V\times\tilde{V}$. If we denote by ϕ the pair $\{u,B\}$ and introduce the operators $\mathfrak{A}\in\mathfrak{L}(\mathfrak{v},\mathfrak{v}')$, $\mathfrak{B}\in\mathfrak{L}(\mathfrak{v}\times\mathfrak{v},\mathfrak{v}')$:

$$\langle\mathfrak{A}\phi_1,\phi_2\rangle=\frac{1}{Re}\int_\Omega\nabla u_1\nabla u_2\,dx+\frac{1}{Rm}\int_\Omega\operatorname{curl}B_1\cdot\operatorname{curl}B_2\,dx,$$

$$\langle\mathfrak{B}(\phi_1,\phi_2),\phi_3\rangle=\int_\Omega\left[(u_1\cdot\nabla)u_2u_3-(B_1\cdot\nabla)B_2u_3+(u_1\cdot\nabla)B_2B_3-(B_1\cdot\nabla)u_2B_3\right]dx;$$

Equations (2.4.1) can be written under the form

$$\frac{d\phi}{dt}+\mathfrak{A}\phi+\mathfrak{B}(\phi,\phi)=\mathfrak{F}.$$

Therefore we can, in the same way as in the previous applications, derive the following regularity and finite dimension results for attractors for M.H.D. equations:

THEOREM 4.1. *Assume that $f\in\mathfrak{C}^\infty(\bar{\Omega})^d$ and then every functional invariant set, X, lies in $\mathfrak{C}^\infty(\bar{\Omega})^{2d}$. Moreover if X is bounded in $H^1(\Omega)^{2d}$, then X is bounded in $H^m(\Omega)^{2d}$ for every $m\geqq 0$.*

THEOREM 4.2. *Every functional invariant set, bounded in $H^1(\Omega)^{2d}$, has finite fractal (and Hausdorff) dimension.*

Remark 4.1. We have considered for the sake of simplicity the homogeneous case (i.e. $\mathfrak{R}=0$). The nonhomogeneous case that could be treated like it has been done for Navier–Stokes equations.

REFERENCES

[1] S. AGMON, *Lectures on Elliptic Boundary Value Problems*, Van Nostrand Mathematical Studies, Van Nostrand, Princeton, NJ, 1965.

[2] T. AUBIN, *Nonlinear Analysis on Manifolds. Monge–Ampere Equations*, Springer-Verlag, Berlin.

[3] A. AVEZ AND Y. BAMBERGER, *Mouvements sphériques des fluides visqueux incompressibles*, J. Mécanique, 17 (1978), pp. 107–145.

[4] A. V. BABIN AND M. I. VISHIK, *Les attracteurs des équations d'évolution aux dérivées partielles et les estimations de leurs dimensions*, Usp. Math. Nauk., 38, 4(232), 1983, pp. 133–187 (in Russian.)

[5] _____, *Attracteurs maximaux dans les équations aux dérivées partielles*, College of France Seminar, Pitman, Boston, 1985.

[6] C. BARDOS AND L. TARTAR, *Sur l'unicité rétrograde des équations paraboliques et quelques questions voisines*, Arch. Rational Mech. Anal., 50 (1973), pp. 10–25.

[7] S. CHANDRASEKHAR, *Hydrodynamic and Hydrodynamic Stability*, Clarendon Press, Oxford, 1961.

[8] P. CONSTANTIN AND C. FOIAS, *Global Lyapunov exponents, Kaplan–Yorke formulas and the dimension of the attractors for* 2D-*Navier–Stokes equations*, Comm. Pure Appl. Math., 38 (1985), pp. 1–37.

[9] P. CONSTANTIN, C. FOIAS, O. MANLEY AND R. TEMAM, *Determining modes and fractal dimension of turbulent flows*, J. Fluid Mech., 150 (1985), pp. 427–440.

[10] P. CONSTANTIN, C. FOIAS AND R. TEMAM, *Attractors representing turbulent flows*, Mem. Amer. Math. Soc., Vol. 53, 1985, #314.

[11] A. DOUADY AND J. OESTERLÉ, *Dimension de Hausdorff des attracteurs*, C. R. Acad. Sci. Paris, 290, Série A (1980), pp. 1135–1138.

[12] D. EBIN AND J. MARSDEN, *Groups of diffeomorphisms and the motion of an incompressible fluid*, Ann. of Math., 92 (1970), pp. 102–163.

[13] C. FOIAS AND G. PRODI, *Sur le comportement global des solutions non stationnaires des équations de Navier–Stokes en dimension* 2, Rend Sem. Mat. Univ. Padova, 39 (1967), pp. 1–34.

[14] C. FOIAS AND R. TEMAM, *Some analytic and geometric properties of the solutions of the evolution Navier–Stokes equations*, J. Math. Pures et Appl., 58 (1979), pp. 339–368.

[15] _____, *On the Hausdorff dimension of an attractor for the two-dimensional Navier–Stokes Equations*, Phys. Letters, 93A (1983), pp. 431–434.

[16] J. M. GHIDAGLIA, *Etude d'écoulements de fluides visqueux incompressibles: comportement pour les grands temps et applications aux attracteurs*, Thèse de Docteur-Ingénieur, Orsay, 1984.

[17] _____, *Régularité des solutions de certains problèmes aux limites liés aux équations d'Euler*, Comm. PDE, 9 (1984), pp. 1237–1264.

[18] _____, *Long time behaviour of solutions of abstract inequalities*, Applications to Thermo-Hydraulic and Magnetohydrodynamic Equations, J. Differential Equations, to appear.

[19] _____, *Some backward uniqueness results*, J. Nonlinear Analysis, T. M. A., to appear.

[20] C. GUILLOPÉ, *Comportement à l'infini des solutions des équations de Navier–Stokes et propriétés des ensembles fonctionnels invariants (ou attracteurs)*, Ann. Inst. Fourier, 32 (1982), pp. 1–37.

[21] S. KOBAYASHI AND K. NOMIZU, *Foundations of Differential Geometry*, Volume I and II, Interscience, New York, 1963.

[22] L. LANDAU AND E. LIFCHITZ, *Electrodynamique des milieux continus*, Editions Mir, Moscou, 1969.

[23] E. LIEB, *On characteristic exponents in turbulence*, Comm. Math. Phys., 92 (1984), pp. 473–480.

[24] J. L. LIONS, *Quelques méthodes de résolution des problèmes aux limites non linéaires*, Dunod, Paris, 1969.

[25] B. MANDELBROT, *Fractal: Form, Chance and Dimension*, W. H. Freeman, San Francisco, 1977.

[26] D. RUELLE, *Large volume limit of distribution of characteristic exponents in turbulence*, Comm. Math. Phys., 87 (1982), pp. 287–302.

[27] _____, *Characteristic exponents for a viscous fluid subjected to time dependent forces*, Comm. Math. Phys., 93 (1984), pp. 285–300.

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

[29] M. SERMANGE AND R. TEMAM, *Some mathematical questions related to the M.H.D. equations*, Comm. Pure Appl. Math., 36 (1983), pp. 634–664.

[30] R. TEMAM, *Behaviour at time $t=0$ of the solutions of semi-linear evolution equations*, J. Differential Equations, 43 (1982), pp. 73–92.

[31] _____, *Navier–Stokes Equations and Nonlinear Functional Analysis*, CBMS Regional Conferences series in Applied Mathematics, 41, Society for Industrial and Applied Mathematics, Philadelphia, 1983.

[32] _____, *Navier–Stokes Equations, Theory and Numerical Analysis*, 3rd ed., North-Holland, Amsterdam, 1984.

[33] _____, *Infinite dimensional dynamical systems in fluid mechanics*, in Nonlinear Functional Analysis and Applications, Univ. California Press, Berkeley, July 1983.

SIAM J. MATH. ANAL.
Vol. 17, No. 5, September 1986

008

EXISTENCE AND CONTROL OF PLASMA EQUILIBRIUM IN A TOKAMAK*

J. BLUM†, T. GALLOUET† AND J. SIMON†

Abstract. The "Tokamak" is a machine where the plasma is confined inside a toroidal vessel by the magnetic field due to currents in external coils. In the present Tokamaks the free plasma boundary is a flux line which is in contact with a given limiter. We first give a simplified model for the plasma axisymmetric equilibrium. We prove, under suitable assumptions on the limiter, the existence of such an equilibrium for given external currents. We then establish the existence of external currents for which the domain aoccupied by the plasma is as close as possible to a given domain. For this problem, which is formulated as an optimal control problem, the first order necessary optimality conditions are obtained by introducing a suitable adjoint state.

Introduction. A model which describes the equilibrium of a plasma in a toroidal machine (a Tokamak) is studied here. The plasma current is obtained by magnetic induction from a primary circuit made of poloidal field coils. In the present Tokamaks the volume occupied by the plasma inside the vacuum vessel is limited by the presence of a "limiter". The configuration is supposed to be axisymmetric, so that the problem can be reduced to a two-dimensional one in the meridian section of the torus.

Section 1 is devoted to the establishment of the model for the plasma axisymmetric equilibrium. Section 2 deals with an annex mathematical problem. In §3 the existence of a solution for the equilibrium problem formulated in §1 is studied. The control of the plasma shape by the external currents is treated in §4.

1. The plasma equilibrium model.

1.1. The Maxwell equations in an axisymmetric configuration. The magnetic induction $\mathbf{B}$ and the current density $\mathbf{j}$ verify the following Maxwell equations:

$$\operatorname{div}\mathbf{B}=0, \tag{1}$$

$$\mathbf{rot}\left(\frac{\mathbf{B}}{\mu}\right)=\mathbf{j}, \tag{2}$$

where μ is the magnetic permeability.

We restrict the problem to the study of axisymmetric configurations, so that in cylindrical coordinates (r,z,φ) the vectors $\mathbf{B}$ and $\mathbf{j}$ are independent of the angle φ.

Equation (1) can then be written

$$\frac{1}{r}\frac{\partial}{\partial r}(rB_r)+\frac{\partial B_z}{\partial z}=0. \tag{3}$$

From (3) one deduces the existence of a scalar function $\psi(r,z)$ called the poloidal flux, such that

$$B_r=-\frac{1}{r}\frac{\partial\psi}{\partial z},\qquad B_z=\frac{1}{r}\frac{\partial\psi}{\partial r}. \tag{4}$$

* Received by the editors November 10, 1983, and in final revised form February 19, 1985.
† Laboratoire d'Analyse Numérique de Paris VI et du CNRS, 4, Place Jussieu, 75230 Paris Cedex 05, France.

From (2) and (4), one can deduce the following partial differential equation for $\psi(r,z)$

$$\mathscr{L}\psi = j_\varphi \tag{5}$$

with

$$\mathscr{L} = -\frac{\partial}{\partial r}\left(\frac{1}{\mu r}\frac{\partial}{\partial r}\right) - \frac{\partial}{\partial z}\left(\frac{1}{\mu r}\frac{\partial}{\partial z}\right)$$

and where j_φ is the toroidal component of the current density (i.e. the φ-component of $\mathbf{j}$).

Let us restrict ourselves to air-transformer Tokamaks, so that μ is constant and equal to the magnetic permeability μ_0 of the vacuum ($\mu_0 = 1$ in Gaussian units). Therefore the operator $\mathscr{L}$ is a linear elliptic operator.

The current density j_φ in the right-hand side of (5) is equal to 0 everywhere except in the coils C_i and in the plasma P.

1.2. Equilibrium conditions for the plasma. In the first Tokamaks (T3, TFR, $\cdots$) the plasma was inside a perfect superconducting shell, on which ψ was constant. The eddy currents on this shell generated the magnetic field necessary to ensure plasma equilibrium (cf. [10], [14]). Simplified models of plasma equilibrium inside a perfect superconducting shell have been studied in [17], [18], [1]. In the present Tokamaks (JET, TFTR, $\cdots$) there is no more superconducting shell and the plasma shape is determined by the currents in the external coils. The plasma is inside a vacuum region V, whose boundary ∂V is the vacuum vessel (see Fig. 1). The plasma boundary is a flux line (ψ = constant). Since an arbitrary constant can be added to ψ according to its definition (4), we can assume that $\psi = 0$ on the plasma boundary. Therefore the plasma domain is defined in the following way

$$P = \{ x \in V \text{ such that } \psi(x) > 0 \}. \tag{6}$$

The plasma particles (electrons and ions) follow the flux lines; they are stopped when they touch the limiter D, which is a piece of metal inside V. This contact condition between the plasma and the limiter can be written

$$P \cap D = \varnothing, \qquad \partial P \cap D \neq 0. \tag{7}$$

In fact the aim of the limiter is to prevent the plasma from touching the vacuum vessel (that is to ensure $\partial P \cap \partial V = \varnothing$).

The plasma is in equilibrium when the kinetic pressure force $\mathbf{grad}\, p$ (p is the plasma kinetic pressure) is equal to the magnetic pressure force ($\mathbf{j} \wedge \mathbf{B}$)

$$\mathbf{grad}\, p = \mathbf{j} \wedge \mathbf{B}. \tag{8}$$

The toroidal plasma current density j_φ can then be deduced from (2), (4) and (8), and it is given by the Grad–Shafranov relation (cf. [7], [13])

$$j_\varphi(r,\psi) = r\frac{\partial p}{\partial \psi} + \frac{1}{2\mu_0 r}\frac{\partial f^2}{\partial \psi} \tag{9}$$

with $f = rB_\varphi$, where B_φ is the toroidal component of $\mathbf{B}$. The functions p and f are the solutions of parabolic diffusion equations, which depend on the shape of the flux lines. These "queer" differential equations (cf. [8]) will not be considered here and, as in [18], we assume, for the sake of simplicity, that j_φ is proportional to ψ in the plasma P. The

proportionality constant λ is unknown here and is determined by the fact that the total plasma current I_p is given (I_p is assumed to be positive). The plasma current density can then be written, with (6),

$$j_\varphi = \lambda \psi^+ 1_V, \qquad I_P = \lambda \int_V \psi^+ \, dx, \tag{10}$$

with $\psi^+ = \sup(\psi, 0)$ and 1_V is the characteristic function of the vacuum region V.

1.3. The equations for the poloidal flux ψ. Physically the problem is set in $R^+ \times R$. But we can restrict it to an open bounded subset Ω of $R^+ \times R$, with a regular boundary $\Gamma = \partial\Omega$, and such that

$$\Omega \subset \{ x = (r,z), r \geqq \underline{r} \} \quad \text{with } \underline{r} > 0.$$

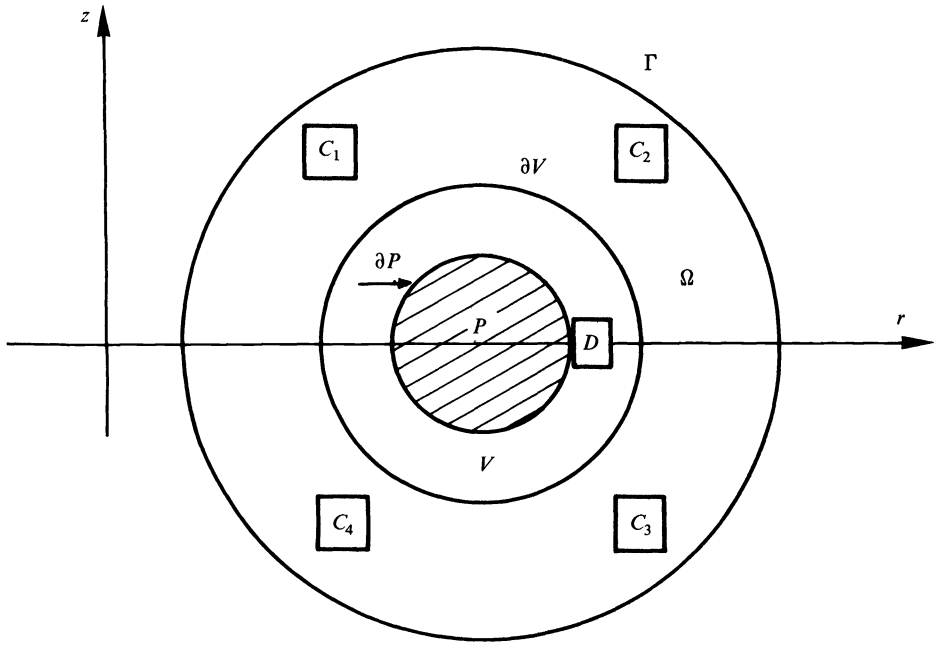


FIG. 1. *The cross section Ω of the torus.*

If Γ is taken sufficiently far from the vacuum vessel and from the coils, then ψ can be assumed to be constant on Γ. Therefore the equations for ψ will be set in such a bounded domain Ω.

The vacuum region V is an open subset of Ω and its boundary ∂V is the vacuum vessel. The sections of the coils $C_1, \cdots, C_k$ are closed subsets of Ω, pairwise disjoint and which do not intersect V. The section D of the limiter is a closed subset of V.

From (5), (7) and (10), one can deduce the equations for the pair (ψ, λ)

$$\begin{aligned} &\mathcal{L}\psi = \lambda \psi^+ 1_V + j \quad \text{in } \Omega, \\ &I_p = \lambda \int_V \psi^+ \, dx, \\ &\psi \text{ is a constant on } \Gamma \end{aligned} \tag{11}$$

and

$$\sup_{x\in D}\psi(x)=0, \tag{12}$$

where j is the external current density, whose support is $\cup_i C_i$. For a continuous function ψ satisfying (11), equation (12) is clearly equivalent to the contact conditions (7), since, by the maximum principle, ψ has no local maximum on D.

Remark 1. From (11) it is clear that, if there exists a zone of positivity of ψ outside V, there is no plasma in this zone.

Remark 2. For a pair (ψ,λ) satisfying (11) and (12), the plasma boundary ∂P may intersect ∂V (see Fig. 2). In this case ∂P is no longer the flux line $\{\psi=0\}$ and this solution is not interesting from a physical point of view. It shows that the limiter D is not correctly located, because in this case it does not prevent the plasma from touching the vacuum vessel ∂V.

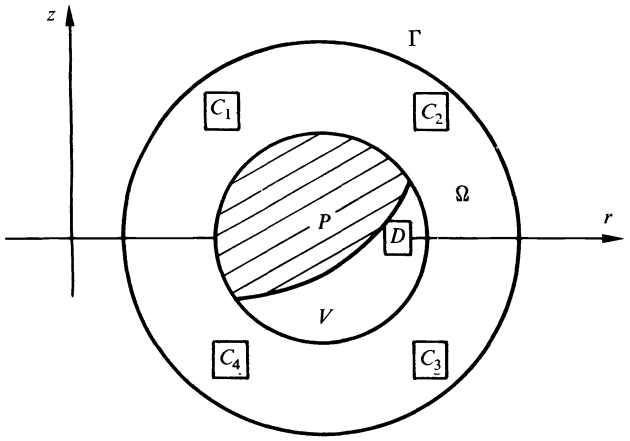


FIG. 2. *A nonphysical solution of equations* (11)–(12).

In §3, we are going to establish under suitable assumptions on D, the existence of (at least) one pair $(\psi,\lambda)\in C^0(\bar{\Omega})\times R$ solution of (11) and (12). In order to prove this, an annex problem will first be studied in §2, namely the existence of a solution of (11), for λ given.

2. An annex problem.

2.1. An existence result. One defines

$$H_c^2(\Omega)=\{v\in H^2(\Omega) \text{ such that } v \text{ is constant on } \Gamma\}.$$

Let us recall that, according to Sobolev's theorem, $H^2(\Omega)\subset C^0(\bar{\Omega})$.

If $\nu>0$, $I>0$ and $f\in L^2(\Omega)$ are given, we are going to study the solutions u of

$$\begin{aligned} &u\in H_c^2(\Omega),\\ &\mathcal{L}u=\nu u^+ 1_V+f \quad \text{in } \Omega,\\ &I=\nu\int_V u^+\,dx. \end{aligned} \tag{13}$$

Many papers have been devoted to the case $f=0$, $V=\Omega$ which corresponds to the plasma equilibrium in a perfect superconducting shell. Particularly existence results have been established by R. Temam [18] and by H. Berestycki–H. Brezis [1] using methods of minimization of suitable functionals and also in [1] using the topological degree of J. Leray–J. Schauder [9].

Following [1] the topological degree method will be used here, because it enables one to prove, in addition to the existence of a solution for each value of the parameters ν, I and f, that there exists a connected set of solutions, when ν is varying. More precisely the following theorem will be proved in §2.6:

THEOREM 1. *For every $I>0$ and $f\in L^2(\Omega)$, there exists a set $\mathscr{C}$ of pairs (u,ν) satisfying* (13) *such that $\mathscr{C}$ is connected in $H_c^2(\Omega)\times R^+$, and such that ν spans $R^+-\{0\}$ when (u,ν) spans $\mathscr{C}$.*

To prove this result, the existence and uniqueness of a solution u of (13) will first be established when ν is sufficiently small, and a priori estimates on the solutions of (13) will be given.

2.2. Uniqueness for ν small. Let $\mu_1, \mu_2 \cdots$ be the sequence of the eigenvalues of the problem

$$\mathscr{L}v=\mu 1_V v, \qquad v\in\mathscr{V}$$

where $\mathscr{V}=\{v\in H_c^2(\Omega)$ such that $\int_\Omega \mathscr{L}v\,dx=0\}$.

LEMMA 1. *If $0<\nu<\mu_2$, equation* (13) *has at most one solution.*

Remark 3. Let $\lambda_1, \lambda_2 \cdots$ be the sequence of the eigenvalues of the problem

$$\mathscr{L}v=\lambda 1_V v, \qquad v\in H_0^1(\Omega).$$

We have: $\mu_1=0$ and $0<\lambda_1<\mu_2\leqq\lambda_2$ (cf. for example [6]).

Let us recall that in the case $f=0$, $V=\Omega$, Lemma 1 is still true with λ_2 instead of μ_2 (cf. [11] and [18]).

Proof of Lemma 1. This result is essentially in H. Berestycki–H. Brezis [1, Thm. 4]. Let us prove it in the following way: let u_1 and u_2 be two distinct solutions of (13). Their difference $w=u_2-u_1$ satisfies

$$\mathscr{L}w=\nu\rho 1_V w, \qquad w\in\mathscr{V} \tag{14}$$

where

$$\rho(x)=\begin{cases}\dfrac{u_2^+(x)-u_1^+(x)}{u_2(x)-u_1(x)} & \text{if } u_2(x)-u_1(x)\neq 0,\\ 0 & \text{if } u_2(x)-u_1(x)=0.\end{cases}$$

Let $\nu_1^\rho, \nu_2^\rho, \cdots$ be the nondecreasing sequence of the eigenvalues of (14). These eigenvalues are nonincreasing as ρ increases. Since $0\leqq\rho\leqq 1$ one has $\nu_2^\rho\geqq\mu_2$. Since $\nu<\mu_2$ one has necessarily $\nu=\nu_1^\rho=0$, which is impossible. □

2.3. A priori estimates.

LEMMA 2. *Every solution u of* (13) *satisfies*

$$|u|_{H^2(\Omega)}\leqq C\left(\nu^2+\frac{1}{\nu}\right)\left(I+\frac{1}{I}\right)\left(1+|f|_{L^2(\Omega)}\right)^2$$

where C only depends on Ω and V.

Since the operator $\mathscr{L}$ is uniformly elliptic with regular coefficients in $\bar{\Omega}$, $\mathscr{L}$ is an isomorphism from $H_0^1(\Omega)\cap H^2(\Omega)$ onto $L^2(\Omega)$. Its inverse operator is denoted K.

For every $s\leqq 0$, there exists $\theta_s\in R$ (depending only on Ω) such that

$$|Kv|_{H^{s+2}(\Omega)}\leqq\theta_s|v|_{H^s(\Omega)},\qquad \forall v\in L^2(\Omega). \tag{15}$$

Proof of Lemma 2. Let u be a solution of (13). Its (constant) value on Γ is denoted u_Γ and we set $P(u)=\{x\in V \text{ such that } u(x)>0\}$. Then we have

$$u=u_\Gamma+\nu K(u^+1_V)+K(f). \tag{16}$$

i) *Upper bound on* $(u-u_\Gamma)$ *in* $L^2(\Omega)$. One has: $|u^+1_V|_{L^1(\Omega)}=I/\nu$; therefore it follows from Sobolev's theorem that

$$|u^+1_V|_{H^{-2}(\Omega)}\leqq C_1\frac{I}{\nu}.$$

Using (15) we deduce

$$|u-u_\Gamma|_{L^2(\Omega)}\leqq C_2\big(I+|f|_{L^2(\Omega)}\big). \tag{17}$$

ii) *Upper bound on* $|u_\Gamma|$. If $u_\Gamma\leqq 0$, we have in $P(u)$

$$2|u_\Gamma|u^+\leqq\big(u^++|u_\Gamma|\big)^2=|u-u_\Gamma|^2,$$

and, integrating on $P(u)$, we deduce

$$2|u_\Gamma|\frac{I}{\nu}\leqq|u-u_\Gamma|^2_{L^2(\Omega)}.$$

If $u_\Gamma\geqq 0$ one has

$$u_\Gamma\leqq|u-u_\Gamma|+u^+ \text{ in } \Omega$$

and, integrating on V, one deduces

$$u_\Gamma|V|\leqq|\Omega|^{1/2}|u-u_\Gamma|_{L^2(\Omega)}+\frac{I}{\nu}.$$

In both cases, using the upper bound on $(u-u_\Gamma)$, one obtains

$$|u_\Gamma|\leqq C_3\Big(\nu+\frac{1}{\nu}\Big)\Big(I+\frac{1}{I}\Big)\big(1+|f|_{L^2(\Omega)}\big)^2.$$

iii) *Upper bound on* u *in* $H^2(\Omega)$. From (15) and (16), one has

$$|u|_{H^2(\Omega)}\leqq|u_\Gamma||\Omega|^{1/2}+\nu\theta_0|u|_{L^2(\Omega)}+\theta_0|f|_{L^2(\Omega)}.$$

The lemma follows from the estimates on $|u-u_\Gamma|_{L^2(\Omega)}$ and $|u_\Gamma|$.

2.4. Transformation of (13). According to the Lax–Milgram theorem, for every $w\in L^2(\Omega)$ there exists a unique $q\in H_c^1(\Omega)$ such that

$$\int_\Omega\left(\frac{1}{r}\nabla q\cdot\nabla v+qv\right)dx=\int_\Omega wv\,dx\quad\forall v\in H_c^1(\Omega),$$

where $H_c^1(\Omega)=\{v\in H^1(\Omega)$ such that v s constant on $\Gamma\}$. One easily verifies that q is the unique solution in $H_c^2(\Omega)$ of

$$\mathscr{L}q+q=w, \qquad \int_\Omega \mathscr{L}q\,dx=0.$$

A linear continuous operator from $L^2(\Omega)$ into $H_c^2(\Omega)$ is then defined by $Lw=q$.

The Lax–Milgram theorem gives also the existence (and the uniqueness) of a solution $\phi\in H_c^1(\Omega)$ of

$$\int_\Omega \left(\frac{1}{r}\nabla\phi\cdot\nabla v+\phi v\right)dx=-\left(I+\int_\Omega f\,dx\right)v_\Gamma \quad \forall v\in H_c^1(\Omega).$$

In fact ϕ is the solution in $H_c^2(\Omega)$ of

$$\mathscr{L}\phi+\phi=0, \qquad \int_\Omega \mathscr{L}\phi\,dx=I+\int_\Omega f\,dx.$$

One defines an operator H from $H_c^2(\Omega)$ into itself by (ν is fixed)

$$H(u)=\phi+L(\nu u^+1_V+f+u).$$

From this definition it is easy to prove the following lemma.

LEMMA 3. *The problem* (13) *is equivalent to*

$$u\in H_c^2(\Omega), \qquad u-H(u)=0. \tag{18}$$

The operator H is compact from $H_c^2(\Omega)$ into itself. Therefore one can define, according to J. Leray–J. Schauder [9, §13, p. 60] the index of the solutions of (18).

Using results of [9], we will prove in §2.5 that, if ν is sufficiently small, there exists a unique solution of (18), with a nonzero index, and we will deduce from this in §2.6 the existence of a connected set $\mathscr{C}$ of solutions (u,ν) of (13) with the properties that ν spans $\mathbb{R}^+-\{0\}$ when (u,ν) spans $\mathscr{C}$.

2.5. Existence of a solution for small ν and computation of its index.

LEMMA 4. *There exists* $\alpha>0$, *which depends only on* Ω, V *and* f, *such that when* $0<\nu\leqq\alpha$, *the problem* (18) *has a unique solution* u_ν, *and the index of* u_ν *is equal to* -1.

This lemma will be proved by transforming continuously, with the help of a parameter t, equation (18) so as to obtain an equation having a unique solution k, whose index is easily computable.

Proof. Let α and ν be two constants such that: $0<\nu\leqq\alpha<\mu_2$.

i) *Transformation of* (18). There exists a unique real number b such that $k=\phi+b$ satisfies

$$\nu\int_V k^+\,dx=I. \tag{19}$$

One defines a function G from $H_c^2(\Omega)\times[0,1]$ into $H_c^2(\Omega)$ by

$$G(v,t)=t\phi+(1-t)k+L\big(\nu v^+1_V+v+tf-(1-t)(\nu k^+1_V+k)\big).$$

Then the equation

$$w\in H_c^2(\Omega), \qquad w-G(w,t)=0 \tag{20}$$

is equivalent to

$$
\begin{aligned}
&w \in H_c^2(\Omega),\\
&\mathscr{L}w = \nu w^+ 1_V + tf + (1-t)(\mathscr{L}k - \nu k^+ 1_V),\\
&\nu \int_V w^+ \, dx = I.
\end{aligned} \tag{21}
$$

For $t=1$, (21) is equivalent to (13) and therefore to (18).

For every t, Lemma 1 shows that there exists at most one solution w. Moreover the operator $v \to G(v,t)$ is compact from $H_c^2(\Omega)$ into itself; therefore one can define the index of the solution w, when this solution exists.

ii) *The case* $t=0$. *Existence of a solution and computation of its index.* As k is a solution of (21), it satisfies (20). We are going to compute its index relatively to this equation.

The condition (19) shows that α can be chosen small enough so that, when $\nu \leqq \alpha$, one has $k>0$ in $\bar{\Omega}$.

Then, in a neighbourhood of k (note that $H^2(\Omega) \hookrightarrow C^0(\bar{\Omega})$), w is positive and $G(w,0)=k+\nu L[(w-k)1_V]+L(w-k)$.

It follows that $G(\cdot,0)$ is Frechet differentiable at the point k, and its derivative is the operator $v \to L(\nu 1_V v + v)$.

According to [9, Conclusion, p. 56], the index of k for equation (20) is equal to the index of 0 for the equation $z - L(\nu 1_V z + z) = 0$ in $H_c^2(\Omega)$, when this latter index exists. This equation is equivalent to

$$
\mathscr{L}z = \nu 1_V z, \qquad \int_\Omega \mathscr{L}z \, dx = 0.
$$

Its eigenvalues are the μ_i defined in §2.2. Since $0=\mu_1<\nu<\mu_2$ and μ_1 is a simple eigenvalue, this index exists and is equal to (-1).

This proves that the index of k for equation (20) is equal to -1.

iii) *Existence of a solution for* $t=1$. The function G: $H_c^2(\Omega)\times[0,1] \to H_c^2(\Omega)$ is continuous, the functions $t \to G(v,t)$ are equicontinuous and, for every t, $v \to G(v,t)$ is compact.

Lemma 2 shows that the possible solutions w of (21), and consequently of (20), are bounded (in $H_c^2(\Omega)$) independently of t.

Then, according to [9], there exist a solution of (20) for $t=1$ and its index is equal to the index of k relatively to (20) for $t=0$.

As $G(\cdot,1)=H$ this proves Lemma 4. □

2.6. Proof of Theorem 1. Let $\underline{\nu}$ and $\bar{\nu}$ be two real numbers such that $0<\underline{\nu} \leqq \alpha < \bar{\nu}$ where α is given by Lemma 4. The operator H defined in §2.4 depending on ν, one defines a function F from $H_c^2(\Omega)\times[\underline{\nu},\bar{\nu}]$ into $H_c^2(\Omega)$ by $F(v,\nu)=H(v)$.

Lemma 3 shows that (13) is equivalent to

$$
u \in H_c^2(\Omega), \qquad u - F(u,\nu) = 0. \tag{22}
$$

Lemma 2 shows that the solutions of (22) are bounded (in $H_c^2(\Omega)$) independently of ν.

Moreover F is continuous, the functions $\nu \to F(v,\nu)$ equicontinuous when v remains in a bounded set, and for every ν the function $v \to F(v,\nu)$ is compact.

Finally for $\nu=\alpha$ the total index of the solutions of (22) is equal to -1 according to Lemma 4.

Theorem 1 of J. Leray–J. Schauder [9] proves then that there exists a set $\mathscr{C}_{\underline{\nu},\bar{\nu}}$ of pairs (u,ν) satisfying (22) such that $\mathscr{C}_{\underline{\nu},\bar{\nu}}$ is connected in $H_c^2(\Omega)\times\mathbb{R}^+$ and such that ν spans $[\underline{\nu},\bar{\nu}]$ when (u,ν) spans $\mathscr{C}_{\underline{\nu},\bar{\nu}}$.

For every positive integer n, setting $\underline{\nu}=\alpha/n$ and $\bar{\nu}=n\alpha$, one obtains a connected set $\mathscr{C}^n$. Since each $\mathscr{C}^n$ contains the point (u_α,α), the reunion $\mathscr{C}$ of all the $\mathscr{C}^n$ is connected, and ν attains in $\mathscr{C}$ all positive values.

This proves Theorem 1. □

2.7. Lower bound of the solutions for small ν.

LEMMA 5. *If* $\nu\leqq I/d(I+|f|_{L^2(\Omega)})$, *then every solution* u *of* (13) *satisfies*

$$u\geqq\frac{I}{\nu|\Omega|}-d'\big(I+|f|_{L^2(\Omega)}\big)\quad \text{in } \bar{\Omega},$$

where d *and* d' *only depend on* Ω.

Proof. One has (recall that $P(u)=\{x\in V,u(x)>0\}$)

$$\int_{P(u)}(u-u_\Gamma)\,dx\leqq\int_\Omega|u-u_\Gamma|\,dx\leqq|\Omega|^{1/2}|u-u_\Gamma|_{L^2(\Omega)},$$

and from (17) it follows that

$$\frac{I}{\nu}-u_\Gamma|P(u)|\leqq d\big(I+|f|_{L^2(\Omega)}\big).$$

If $I/\nu\geqq d(I+|f|_{L^2(\Omega)})$ one has $u_\Gamma\geqq 0$ and, since $|P(u)|\leqq|\Omega|$, it comes

$$u_\Gamma\geqq\frac{I}{\nu|\Omega|}-\frac{d}{|\Omega|}\big(I+|f|_{L^2(\Omega)}\big).$$

The maximum principle shows that $K(u^+1_V)\geqq 0$ where the operator K has been defined in §2.3. It follows from (15) that $|K(f)|_{L^\infty(\Omega)}\leqq d_1|f|_{L^2(\Omega)}$ and, with (16) one has

$$u\geqq\frac{I}{\nu|\Omega|}-\frac{d}{|\Omega|}\big(I+|f|_{L^2(\Omega)}\big)-d_1|f|_{L^2(\Omega)}.$$ □

2.8. An estimate on $P(u)$ for large ν.

LEMMA 6. *Assume that* $\mathrm{Supp}(f)\cap V=\varnothing$. *Then there exists* C *depending only on* Ω *such that for every* $\nu>0$ *and* u *solution of* (13) *one has* $B(x_0,C/\sqrt{\nu})\not\subset P(u)$, *for every* $x_0\in\Omega$ (*where* $B(x_0,R)$ *designates the ball centered on* x_0 *and of radius* R).

Proof. Let us suppose that there exists a ball B, of radius R, included in $P(u)$. Then

$$\mathscr{L}u=\nu u \quad \text{in } B,$$
$$u\geqq 0 \quad \text{on } \partial B.$$

Let $\lambda_1(B)$ be the first eigenvalue of

$$\mathscr{L}\varphi=\lambda\varphi,\qquad \varphi\in H_0^1(B)$$

and let φ_1 be a related positive eigenfunction. Then,

$$\nu\int_B u\varphi_1\,dx=\int_B \mathscr{L}u\varphi_1\,dx=\int_B u\mathscr{L}\varphi_1\,dx+\int_{\partial B}\frac{u}{r}\frac{\partial\varphi_1}{\partial n}\,d\sigma$$

$$=\lambda_1(B)\int_B u\varphi_1\,dx+\int_{\partial B}\frac{u}{r}\frac{\partial\varphi_1}{\partial n}\,d\sigma,$$

whence

$$(\nu-\lambda_1(B))\int_B u\varphi_1\,dx=\int_{\partial B}\frac{u}{r}\frac{\partial\varphi_1}{\partial n}\,d\sigma\leqq 0.$$

This implies $\nu\leqq\lambda_1(B)$.

Since $\lambda_1(B)<C^2/R^2$ (where C is a constant depending only on Ω), Lemma 6 is proved. □

3. Existence of solutions for (11)–(12).

3.1. An existence result. The aim of this section is to find a solution (ψ,λ) of equations (11)–(12). The total plasma current I_p and the external current density j are given such that

$$\text{(23)}\qquad \begin{aligned} &I_p>0,\\ &j\in L^2(\Omega),\quad \operatorname{supp} j\cap V=\varnothing. \end{aligned}$$

Let us set, in (13), $f=j$ and $I=I_p$. According to Theorem 1, one can define a unique set C^* by

> C^* is the greatest set of pairs $(u,\nu)\in H_c^2(\Omega)\times\mathbb{R}^+$ satisfying (13), such that C^* is connected in $H_c^2(\Omega)\times\mathbb{R}^+$, and such that ν spans $\mathbb{R}^+-\{0\}$ when (u,ν) spans C^*.

Let us recall from Lemma 1 that if $0<\nu<\mu_2$ there is a unique solution $u=u_\nu$ of (13). Then $(u_\nu,\nu)\in C^*$.

THEOREM 2. *Assume* (23) *and*

$$\text{(24)}\qquad \textit{there exists } (v,\gamma)\in C^* \textit{ such that } v\leqq 0 \textit{ in } D.$$

Then there exists a pair $(\psi,\lambda)\in H_c^2(\Omega)\times\mathbb{R}^+$, *solution of* (11)–(12).

Proof of Theorem 2. Since $u\to\sup_{x\in D}u(x)$ is continuous from $C^0(\bar{\Omega})$ (and therefore from $H_c^2(\Omega)$) into $\mathbb{R}$, the set

$$\left\{\sup_{x\in D}u(x),(u,\nu)\in C^*\right\}\text{ is connected in }\mathbb{R}.$$

Lemma 5 shows that there exists β, $0<\beta<\mu_2$, such that $\sup_{x\in D}u(x)\geqq 0$. By hypothesis (24) one has $\sup_{x\in D}v(x)\leqq 0$. Then there exists $(\psi,\lambda)\in C^*$ with $\sup_{x\in D}\psi(x)=0$. The pair (ψ,λ) satisfies (11)–(12). □

Remark 4. Condition (24) is hard to verify in practice. Letting (v,γ) be a solution of (13), it is an open problem to know if $(v,\gamma)\in C^*$, except if $\gamma<\mu_2$. Indeed in this latter case, from the uniqueness of Lemma 1, one has $(v,\gamma)\in C^*$.

Remark 5. Every pair $(u,\nu)\in C^*$ defines an admissible zone for the limiter D. Indeed if $D\subset V\setminus P(u)$, there exists a solution of (11)–(12). This zone is not empty as

soon as $\nu > \lambda'_1$ where λ'_1 is the first eigenvalue of the Dirichlet problem for the operator $\mathscr{L}$ in$V$. (Note that $\lambda'_1 < \mu_2$).

Another possibility to verify (24) is to study the behaviour of $P(u)$, for (u,ν) solution of (13), as $\nu \to +\infty$. This is done in the next sections 3.2 and 3.3.

3.2. Existence by infinitesimal change of the limiter. For two closed subsets E, F of R^2 we recall the definition of the Hausdorff distance from E to F

$$d_H(E,F) = \mathrm{Sup}\left\{ \sup_{x\in E} d(x,F);\ \sup_{x\in F} d(x,E) \right\}.$$

THEOREM 3. *Let $\tilde{D}$ be a closed subset of V, and $\varepsilon > 0$. Then there exists a closed subset D of V such that*

i) $d_h(\tilde{D}, D) \leqq \varepsilon$.

ii) *There exists* $(\psi,\lambda) \in H_c^2(\Omega)\times\mathbb{R}^+$, *solution of* (11)–(12).

Proof of Theorem 3. We can assume that $\varepsilon < d(\tilde{D}, \Omega\setminus V)$. From Theorem 1, there exists $(u,\nu) \in C^*$ such that $\nu = C^2/\varepsilon^2$, where C is defined by Lemma 6. From Lemma 6 one has $B(x,\varepsilon) \not\subset P(u)$ for every $x \in \tilde{D}$. Set $D = \overline{U_{x\in\tilde{D}} B(x,\varepsilon)} \setminus P(u)$. One has $D \subset V$, $d_H(D,\tilde{D}) \leqq \varepsilon$, D closed. Moreover, $u \leqq 0$ on D; then from Theorem 2 we deduce ii). □

Remark 6. The case of punctual limiters.

Let $\omega = \{ d \in V$ such that there exists $(\psi,\lambda) \in H_c^2(\Omega)\times\mathbb{R}^+$ solution of (11)–(12) with $D = \{d\}\}$. Then ω is dense in V.

Indeed in Theorem 3 if $\tilde{D}$ is reduced to a single point, then in the proof D can be chosen as a single point.

Remark 7. The cylindrical case. If the machine is cylindrical instead of toroidal, the operator $\mathscr{L}$ is replaced by $-\Delta$. In this case if Ω is a ball centered at 0 and if $V = \Omega$, $j = 0$, then there exists a solution of (11)–(12) iff $0 \notin D$. then the set ω in Remark 6 is $\Omega\setminus\{0\}$.

An example for which the existence holds only for some convenient values of the inducing current j in one single coil C_1 is given in J. Simon [15].

3.3. Open questions. In order to prove condition (24) in Theorem 2 it would be interesting to know the answer to the following questions.

Open question 1. Is the set of pairs (u,ν) satisfying (13) connected in $C^0(\bar{\Omega})\times\mathbb{R}^+$?

On the other hand, we can define "variational solutions" of (13) as in [18]. Indeed, let us set for $v \in H_c^1(\Omega)$

$$E(v) = \int_\Omega \left(\frac{1}{r} |\nabla v|^2 - \nu |v^+|^2 1_V - 2jv \right) dx + 2\left(I_p + \int_\Omega j\,dx \right) v_\Gamma.$$

By using a simple adaptation of the method of R. Temam [18] one can prove that for every $\nu > 0$, E reaches its minimum on the set $K = \{ v \in H_c^1(\Omega),\ \int_V v^+\,dx = I_p/\nu \}$. Moreover each minimizing point of E on K is a solution, called "variational solution", of (13).

By an adaptation of the method of L. Caffarelli–A. Friedman [4] (where $V = \Omega$, $j = 0$) one may prove that, for every variational solution u, $\dim P(u) \to 0$ as $\nu \to \infty$, and $P(u)$ is asymptotically located on the set $\partial V^+ = \{ x \in \bar{V}, r(x) \geqq r(y)\ \forall y \in V \}$.

Therefore condition (24) is satisfied if one can answer positively to the following question.

Open question 2. Is the set of pairs (u,ν) such that u is a variational solution connected in $C^0(\bar{\Omega})\times\mathbb{R}^+$?

At last for some particular geometries and currents it is perhaps also possible to answer positively to the next question.

Open question 3. For ν large enough does (13) have a unique solution? This is not always the case. Indeed the counterexample given by D. G. Schaeffer [12] relative to the operator $-\Delta$, can be adapted to the operator $\mathscr{L}$.

3.4. Computational aspects. More sophisticated models can be written for the axisymmetric equilibrium in a Tokamak, including a more general law for the plasma current density and a nonlinear operator $\mathscr{L}$; indeed, in an iron-transformer Tokamak, the nonlinearity of $\mathscr{L}$ is due to the fact that the constant μ_0 is replaced by a function $\mu(|\nabla\psi|)$. These models are numerically solved in [2] by a finite element method and by iterative algorithms for the nonlinearities; they are applied to the simulation of the axisymmetric equilibrium configurations in TFR 600 (Tokamak of Fontenay-aux-Roses) and JET (Joint European Torus).

4. Control of the plasma shape.

4.1. Formulation of the problem. Let us assume that the current density is homogeneous in each coil C_i and let I_i be the total current flowing in C_i; then j can be written

$$j=\sum_{i=1}^{k}\frac{I_i}{|C_i|}1_{C_i},$$

where $|C_i|$ denotes the Lebesgue measure of C_i and k the number of coils. One intends to determine the currents I_i in the coils C_i, so that the plasma boundary ∂P is as close as possible to a desired boundary Σ. In order to achieve this, we want to realize "at best" $\psi=0$ on Σ. Therefore we shall minimize the following cost-function

$$J(\psi)=\int_{\Sigma}\psi^2 d\sigma$$

on the set Ψ of the solutions of (11)–(12) when the currents I_i are varying, i.e.,

$$\Psi=\left\{\psi\in C^0(\bar{\Omega}),\ \text{such that } \exists\lambda>0,\ I_1,\cdots,I_k,\ \text{so that } (\psi,\lambda) \text{ satisfy (11)–(12)}\right\}.$$

Let us mention that the total plasma current I_p is fixed. We are looking for a solution $\psi\in\Psi$ of the following optimal control problem

$$J(\psi)=\inf_{\tilde{\psi}\in\Psi}J(\tilde{\psi}). \tag{25}$$

Notice that since there is only a finite number k of control parameters, the minimum of J will generally be different from 0 and ∂P will not be exactly identical to Σ. It is reasonable to assume that Σ is in contact with the limiter D, because the plasma boundary must satisfy this condition. But this hypothesis will not be used in the following. On the other hand we shall use the following hypothesis

$$\Omega-\bigcup_{i=1}^{k}C_i \text{ is connected; } \Sigma \text{ is the boundary of a regular open subset of } V. \tag{26}$$

4.2. An a priori estimate.

LEMMA 7. *Every solution* $\psi\in C^0(\bar{\Omega})$ *of* (11) *satisfies*

$$\sum_{i=1}^{k}|I_i|+|\psi_\Gamma|\leqq e\left(I_p+\sqrt{j(\psi)}\right)$$

where the constant e only depends on Ω, Σ, $C_1,\cdots,C_k$.

Proof of Lemma 7. From (16) it follows that $\psi=\psi_\Gamma+K(\lambda\psi^+1_V)+K(j)$. One has $|\lambda\psi^+1_V|_{L^1(\Omega)}=I_p$. Since $L^1(\Omega)\hookrightarrow H^{-s}(\Omega)$, for all $s>1$, it follows from (15) that, taking some $s\in\left]1,\frac{3}{2}\right[$, one has

$$|K(\lambda\psi^+1_V)|_{H^{-s+2}(\Omega)}\leqq e_1 I_p.$$

Since $|\psi|_{L^2(\Sigma)}=\sqrt{j(\psi)}$, and since the trace operator is linear continuous from $H^2(\Omega)$ into $L^2(\Sigma)$, one deduces

$$|K(j)+\psi_\Gamma|_{L^2(\Sigma)}\leqq e_2 I_p+\sqrt{j(\psi)}\,.$$

Let $\bar v$ be the mean value of the function v on Σ, i.e. $\bar v=(1/|\Sigma|)\int_\Sigma v\,d\sigma$. One has $|\bar v|_{L^2(\Sigma)}\leqq|v|_{L^2(\Sigma)}$. Let us set $g=K(j)-\overline{K(j)}$. Then

$$|g|_{L^2(\Sigma)}\leqq 2|K(j)+\psi_\Gamma|_{L^2(\Sigma)}\leqq 2\left(e_2 I_p+\sqrt{J(\psi)}\right).$$

i) We shall now prove that there exists $e_3>0$ such that

$$|g|_{L^2(\Sigma)}\geqq e_3\sum_i|I_i|.$$

Let us suppose that $|g|_{L^2(\Sigma)}=0$, i.e. $g=0$ on Σ. One has $\mathscr{L}g=j$ which is zero in $\Omega-\bigcup_i C_i$. Then, with (26), we deduce that $g=0$ on $\Omega-\bigcup_i C_i$.

Then for each i, one has

$$I_i=\int_{C_i} j\,dx=\int_{C_i}\mathscr{L}g\,dx=-\int_{\partial C_i}\frac{1}{r}\frac{\partial g}{\partial n}\,d\sigma=0.$$

Then the function $(I_1,\cdots,I_k)\to|g|_{L^2(\Sigma)}$ is a norm on $\mathbb{R}^k$. This proves the existence of the constant e_3.

ii) It remains to bound ψ_Γ. We have

$$|\psi_\Gamma|\leqq|\overline{K(j)}+\psi_\Gamma|+|\overline{K(j)}|$$

and

$$|\overline{K(j)}+\psi_\Gamma|\leqq|\Sigma|^{-1/2}|K(j)+\psi_\Gamma|_{L^2(\Sigma)},$$

$$|\overline{K(j)}|=\left|\sum_i\frac{I_i}{|C_i|}\overline{K(1_{C_i})}\right|\leqq e_4\sum_i|I_i|,$$

whence

$$|\psi_\Gamma|\leqq\left(|\Sigma|^{-1/2}+\frac{2e_4}{e_3}\right)|K(j)+\psi_\Gamma|_{L^2(\Sigma)}\leqq e_5\left(I_p+\sqrt{J(\psi)}\right).$$

4.3. Existence of an optimal control. We assume here that λ remains bounded when j and ψ_Γ are bounded. More precisely

(27) For every $a>0$ and $\psi\in\Psi$ such that $\sum_i|I_i|+|\psi_\Gamma|\leqq a$ one has $\lambda\leqq A$,

where A depends on a, Ω, V, I_p and on the C_i.

Remark 8. We can show that (27) is satisfied if equation (13) has only variational solutions, and in particular if, for every ν and f there exists a unique solution of (13).

THEOREM 4. *Assume that* $\Psi \neq \varnothing$ *and that* (26) *and* (27) *are verified. Then there exists a solution* ψ_0 *of* (25).

Proof. Let $\{\psi^n\}$ be a minimizing sequence of J in Ψ, i.e.,

$$J(\psi^n) \to \inf_{\psi \in \Psi} J(\psi) \quad \text{as } n \to \infty.$$

Let λ^n and $I_1^n, \cdots, I_k^n$ be the proportionality coefficient and the currents relative to ψ^n.

Lemma 7 shows that I_i^n and ψ_Γ^n are bounded independently of n.

Then j^n is bounded in $L^2(\Omega)$ and Lemma 5 shows that there exists $b>0$, independent of n, such that $\lambda^n \geqq b$ (otherwise ψ_Γ^n would not be bounded).

Moreover assumption (27) shows that $\lambda^n \leqq B$. Then Lemma 2 shows that the ψ^n are bounded in $H^2(\Omega)$.

The imbedding of $H^2(\Omega)$ into $C^0(\bar{\Omega})$ being compact, one can extract a subsequence, still denoted ψ^n such that as $n \to +\infty$

$$\psi^n \to \psi_0 \quad \text{in } C^0(\bar{\Omega}),$$
$$\lambda^n \to \lambda_0, \qquad I_i^n \to I_i^0.$$

Passing to the limit in the equations (11) and (12) relative to ψ^n, one shows that $\psi_0 \in \Psi$. Moreover $J(\psi^n) \to J(\psi_0)$, therefore ψ_0 verifies (25).

Remark 9. *Control of the variational solutions*. Let Ψ_{var} be the subset of Ψ constituted by the variational solutions. Assume that $\Psi_{\text{var}} \neq \varnothing$ and (26) is satisfied. Then there exists $\psi_0 \in \Psi_{\text{var}}$ such that

$$J(\psi_0) = \inf_{\psi \in \Psi_{\text{var}}} J(\psi).$$

This is a consequence of Remark 8, and of the fact that a limit of variational solutions is itself a variational solution.

4.4. Differentiation of a solution with respect to the currents in the coils. If (ψ, λ) is a solution of (11)–(12) relative to $j = \sum_i I_i (1_{C_i}/|C_i|)$, we intend to find, in a neighbourhood of (ψ, λ), a solution which depends continuously on the currents.

For this it is enough to find a solution which depends continuously on each current. Let us consider the variation with respect to I_1.

We assume (this assumption may be weakened as we shall see later) that

(28) D is reduced to a point d.

For every $t \in R$, one looks for a solution of (11) and (12) relative to $j_t = j + t 1_{C_1}/|C_1|$, i.e. for a solution of

$$\begin{aligned} &\psi_t \in H_c^2(\Omega), \qquad \lambda_t > 0, \\ &\mathcal{L}\psi_t - \lambda_t \psi_t^+ 1_V = j_t, \\ &\int_\Omega \mathcal{L}\psi_t \, dx = I_p + \int_\Omega j_t \, dx, \\ &\psi_t(d) = 0, \end{aligned} \tag{29}$$

such that $t \to (\psi_t, \lambda_t)$ is differentiable. (With $(\psi_0, \lambda_0) = (\psi, \lambda)$.)

First let us differentiate formally this equation so as to obtain the equations satisfied by the derivative $\psi' = \partial\psi_t/\partial t$, $\lambda' = \partial\lambda_t/\partial t$ at the point $t=0$. We set $P = \{x \in V, \psi(x) > 0\}$. One obtains

$$
\begin{aligned}
&\psi' \in H_c^2(\Omega), \qquad \lambda' \in \mathbb{R}, \\
&\mathscr{L}\psi' - \lambda 1_P \psi' - \lambda' \psi^+ 1_V = \frac{1_{C_1}}{|C_1|}, \\
&\int_\Omega \mathscr{L}\psi' \, dx = 1, \\
&\psi'(d) = 0.
\end{aligned} \tag{30}
$$

We are going to prove this "formal" differentiation by using the implicit function theorem. To do this we introduce the eigenvalue problem

$$
\begin{aligned}
&\varphi \in \mathscr{V} = \left\{ v \in H_c^2(\Omega), \int_\Omega \mathscr{L} v \, dx = 0 \right\}, \\
&\mathscr{L}\varphi = \nu 1_P \varphi.
\end{aligned} \tag{31}
$$

If λ is not an eigenvalue of (31), one defines the unique function φ_1 by

$$
\varphi_1 \in \mathscr{V}, \qquad \mathscr{L}\varphi_1 - \lambda 1_P \varphi_1 = \psi^+ 1_V. \tag{32}
$$

THEOREM 5. *The pair* $(\psi, \lambda) \in H_c^2(\Omega) \times R$ *being a solution of* (11)–(12), *one assumes that*

$$
\nabla\psi(x) \neq 0 \text{ for every } x \in \bar{V} \text{ such that } \psi(x) = 0, \tag{33}
$$

$$
\lambda \text{ is not an eigenvalue of } (31) \text{ and } \varphi_1(d) \neq 0. \tag{34}
$$

Then, for every small enough t, *there exists a unique solution of* (29) *such that* $t \to (\psi_t, \lambda_t)$ *is continuously differentiable into* $H_c^2(\Omega) \times \mathbb{R}$ *and such that* $(\psi_0, \lambda_0) = (\psi, \lambda)$.

The derivative for $t=0$ *is the unique solution* (ψ', λ') *of* (30).

Remarks 10.

i) From (29) one can deduce that $\psi \in C^1(\bar{\Omega})$ and (33) is therefore meaningful.

ii) Let us assume that $\lambda < \mu_2$, where μ_2 is defined in §2.2. Then λ is not an eigenvalue of (31).

Indeed, as it has been seen in the proof of Lemma 1, the first eigenvalue of (31) is zero, and the second is larger than μ_2.

Remark. Analogous results of derivability have been obtained by A. Dervieux [5], for the equilibrium problem without inductive currents, vacuum vessel nor limiter.

Proof of Theorem 5. Let $p > 2$ be given and let us set $X = \{v \in W_c^{2,p}(\Omega), v(d) = 0\}$. Provided with the norm of $W^{2,p}(\Omega)$, X is a Banach space included in $C^1(\bar{\Omega})$.

One defines F: $R \times X \times R \to L^p(\Omega) \times \mathbb{R}$ by

$$
F(t; v, b) = \left(\mathscr{L} v - b v^+ 1_V - j_t, \int_\Omega \mathscr{L} v \, dx - I_p - \int_\Omega j_t \, dx \right).
$$

One looks for a solution of $F(t; \psi_t, \lambda_t) = 0$ in the neighbourhood of the solution (ψ, λ) relative to $t=0$. The implicit function theorem gives the announced results, provided

that one verifies the following properties:

(35) There exists a neighbourhood $\mathcal{V}'$ of ψ in X such that $F \in C^1(\mathbb{R} \times \mathcal{V}' \times \mathbb{R}, L^p(\Omega) \times \mathbb{R})$.

(36) The derivative $A = (\partial F/\partial(v,b))(0;\psi,\lambda)$ is an isomorphism from $X \times \mathbb{R}$ onto $L^p(\Omega) \times \mathbb{R}$.

i) *Proof of* (35). The only difficulty comes from the nonlinear term $v^+ 1_V$. The function $v \to v^+ 1_V$ is differentiable from $W^{2,p}(\Omega)$ into $L^p(\Omega)$ at any point w such that the measure of $\{x \in V, w(x) = 0\}$ is zero, and its derivative is the function $G(w)$: $v \to 1_{P_w} v$ where $P_w = \{x \in V, w(x) > 0\}$ (cf. for example [6] for the proof of these points).

To prove (35), we have to verify that:

(37) for every $w \in \mathcal{V}'$ the measure of the set $\{x \in V, w(x) = 0\}$ is zero,

(38) $G \in C^0(\mathcal{V}', \mathcal{L}(W^{2,p}(\Omega); L^p(\Omega)))$.

From the assumption (33) it follows that there exists a neighbourhood of ψ in $C^1(\bar{\Omega})$, and therefore a neighbourhood $\mathcal{V}'$ of ψ in X such that $\forall w \in \mathcal{V}'$, one has

$$\nabla w(x) \neq 0 \text{ for any } x \in \bar{V} \text{ such that } w(x) = 0.$$

This yields (37) since, according to G. Stampacchia [16], one has

$$1_{w=0} \nabla w = 0 \quad \text{a.e. in } V.$$

Moreover, when $w_n \to w$ in $\mathcal{V}'$ one has, since the measure of $\{x \in V, w(x) = 0\}$ is zero, $1_{P_{w_n}} \to 1_{P_w}$ in $L^1(\Omega)$. Therefore, for any $v \in W^{2,p}(\Omega)$, one has

$$\left| (1_{P_{w_n}} - 1_{P_w}) v \right|_{L^p(\Omega)} \leqq \left| 1_{P_{w_n}} - 1_{P_w} \right|^{1/p}_{L^1(\Omega)} |v|_{L^\infty(\Omega)}$$

which yields (38).

ii) *Proof of* (36). The operator $A \in \mathcal{L}(X \times \mathbb{R}; L^p(\Omega) \times \mathbb{R})$ is defined by

$$A(\varphi, \mu) = \left(\mathcal{L}\varphi - \lambda 1_P \varphi - \mu \psi^+ 1_V, \int_\Omega \mathcal{L}\varphi \, dx \right).$$

One has to show that, for every $(g,a) \in L^p(\Omega) \times \mathbb{R}$, there exists a unique solution (φ, μ) of

$$\begin{aligned} &\varphi \in W_c^{2,p}(\Omega), \qquad \mu \in \mathbb{R}, \\ &\mathcal{L}\varphi - \lambda 1_P \varphi - \mu \psi^+ 1_V = g, \\ &\int_\Omega \mathcal{L}\varphi \, dx = a, \\ &\varphi(d) = 0. \end{aligned} \tag{39}$$

Since λ is not an eigenvalue of (31) there exists a unique function φ_2 such that

$$\begin{aligned} &\varphi_2 \in H_c^2(\Omega), \\ &\mathcal{L}\varphi_2 - \lambda 1_P \varphi_2 = g, \\ &\int_\Omega \mathcal{L}\varphi_2 \, dx = a. \end{aligned}$$

It is clear that φ_1 and φ_2 belong to $W_c^{2,p}(\Omega)$; therefore (φ,μ) is a solution of (39) if and only if

$$\varphi=\mu\varphi_1+\varphi_2, \qquad \varphi(d)=0. \tag{40}$$

Since one has supposed $\varphi_1(d)\neq 0$, (40) (and therefore (39)) has a unique solution (φ,μ). This proves (36). □

Remark 11. In Theorem 5, we have assumed that D is reduced to one point. This assumption is not necessary. In fact if D is constituted by a finite number of points and if we assume that the maximum of ψ on D is reached at a unique point, it is easy to see that Theorem 5 is still true.

More generally let us suppose that D is a closed subset of V, with a regular boundary and that the maximum of ψ on D is reached at a unique point d. Then one can prove a result, which is similar to that of Theorem 5. By using additional assumptions, in particular that the curvatures of the boundaries of P and D at the point d are different, one can show that ψ_t reaches its maximum at a point d_t which depends regularly on t. The derivative of $t\to\psi_t$ for $t=0$ is still given by (30).

4.5. Optimality conditions. In this paragraph we still assume that D is reduced to a point d (this hypothesis can be weakened as we have seen in the remark at §4.4).

Let ψ_0 be an optimal control, i.e. a solution of (25). Thus there exists a unique $\lambda_0>0$ and unique currents $I_1^0,\cdots,I_k^0$ such that (ψ_0,λ_0) satisfies (11) and (12).

One denotes by δ_d the Dirac measure at point d, and δ_Σ the measure on Σ defined by

$$(\delta_\Sigma,v)=\int_\Sigma v\,d\sigma \quad \forall v\in C^0(\bar\Omega).$$

(σ is the 1-dimensional Lebesgue measure on Σ). We set $P_0=\{x\in V,\psi_0(x)>0\}$.

THEOREM 6. *We assume that* (ψ_0,λ_0) *satisfies the assumptions* (33) *and* (34). *Then there exists a unique solution* q *of*

$$\begin{aligned} &q\in W_c^{1,p'}(\Omega) \quad \textit{where } 1<p'<2,\\ &\mathcal{L}q-\lambda_0 1_{P_0}q=\psi_0\delta_\Sigma-\left(\lambda_0\int_{P_0}q\,dx+\int_\Sigma\psi_0\,d\sigma\right)\delta_d,\\ &\int_V q\psi_0^+\,dx=0, \end{aligned} \tag{41}$$

and one has the following necessary optimality condition

$$\frac{1}{|C_i|}\int_{C_i}q\,dx=q_\Gamma \quad \forall i.$$

Proof of Theorem 6 .

i) *Differentiation of the cost function.* Theorem 5 enables to define in a unique way, for t small enough, a solution (ψ_t,λ_t) of (11)–(12) relative to $j_t=j_0+t1_{C_1}/|C_1|$.

Since ψ_0 is optimal one has

$$\int_\Sigma\psi_t^2\,d\sigma\geqq\int_\Sigma\psi_0^2\,d\sigma \quad \forall t.$$

Since $t \to \psi_t$ is differentiable into $H_c^2(\Omega)$, one has

$$\int_\Sigma \psi_0 \psi' \, d\sigma = 0 \tag{42}$$

where (ψ', λ') is the unique solution of (30) (relative to $(\psi, \lambda) = (\psi_0, \lambda_0)$).

By introducing an appropriate adjoint state, we are going to transform this condition (42).

ii) *Definition of the adjoint state q.* As it has been mentioned in the part ii) of the proof of Theorem 5, the operator A which is defined by

$$A(\varphi, \mu) = \left(\mathscr{L}\varphi - \lambda_0 1_{P_0} \varphi - \mu \psi_0^+ 1_V, \int_\Omega \mathscr{L}\varphi \, dx \right)$$

is linear continuous from $X \times \mathbb{R}$ onto $L^p(\Omega) \times \mathbb{R}$. Let us recall that $X = \{ v \in W_c^{2,p}(\Omega), v(d) = 0 \}$.

The adjoint operator A^* of A is therefore linear continuous from $L^{p'}(\Omega) \times \mathbb{R}$ onto $X' \times \mathbb{R}$ with $p' = p/(p-1)$.

As the embedding of X into $C^0(\bar{\Omega})$ is continuous, the measure $\psi_0 \delta_\Sigma$ is an element of X'. Then one defines (q, h) in a unique way by $A'(q, h) = (\psi_0 \delta_\Sigma, 0)$, i.e. by

$$\begin{aligned} &q \in L^{p'}(\Omega), \qquad h \in \mathbb{R}, \\ &\int_\Omega q \left(\mathscr{L}\varphi - \lambda_0 1_{P_0} \varphi - \mu \psi_0^+ 1_V \right) dx + h \int_\Omega \mathscr{L}\varphi \, dx = \int_\Sigma \psi_0 \varphi \, d\sigma \quad \forall \varphi \in X, \ \mu \in \mathbb{R}. \end{aligned} \tag{43}$$

When v spans $W_c^{2,p}(\Omega)$, $v - v(d)$ spans X and then (43) is equivalent to

$$\begin{aligned} &q \in L^{p'}(\Omega), \qquad h \in \mathbb{R}, \\ &\int_\Omega q \left(\mathscr{L}v - \lambda_0 1_{P_0} v \right) dx + h \int_\Omega \mathscr{L}v \, dx = \int_\Sigma \psi_0 v \, d\sigma - v(d) \left[\lambda_0 \int_{P_0} q \, dx + \int_\Sigma \psi_0 \, d\sigma \right], \\ &\int_V q \psi_0^+ \, dx = 0 \quad \forall v \in W_c^{2,p}(\Omega). \end{aligned} \tag{44}$$

iii) *Characterization* (41) *of q.* Let us set

$$\mu = \lambda_0 1_{P_0} q + \psi_0 \delta_\Sigma - \left(\lambda_0 \int_{P_0} q \, dx + \int_\Sigma \psi_0 \, d\sigma \right) \delta_d.$$

One has $\mu \in W^{-1,p'}(\Omega)$ since the measures belong to this space.

Let us consider the solution η of

$$\eta \in W_0^{1,p'}(\Omega), \qquad \mathscr{L}\eta = \mu, \tag{45}$$

and let us assume for a while that η is the unique solution of

$$\eta \in L^{p'}(\Omega), \quad \int_\Omega \eta \mathscr{L}v \, dx = (\mu, v) \quad \forall v \in W^{2,p}(\Omega) \cap W_0^{1,p}(\Omega). \tag{46}$$

Since $q + h$ satisfies (46), one has $q + h = \eta$, which establishes (41).

Reciprocally if q satisfies (41) one has $q - q_\Gamma = \eta$, then for $h = -q_\Gamma$ the pair (q, h) satisfies (44), therefore (41) has a unique solution.

iv) *Equivalence of* (45) *and* (46). Let us first prove that η satisfies (46). One sets $v = v_1 + v_2$ where v_1 equals zero in a neighbourhood of Γ and v_2 equals zero in a neighbourhood of $\Sigma \cup \{d\}$. If $v_1 \in \mathcal{D}(\Omega)$ one has

$$\int_\Omega \eta \mathcal{L} v_1 \, dx = (\mathcal{L}\eta, v_1)_{\mathcal{D}'(\Omega)\times\mathcal{D}(\Omega)} = (\mu, v_1)$$

and by density one obtains this same result for every $v_1 \in W_c^{2,p}(\Omega)$, such that v_1 equals zero in a neighbourhood of Γ.

Moreover, as $\mu = \lambda_0 1_{p_0} q$ in the support S of v_2 one has $\eta \in W^{2,p'}(S)$ and

$$\int_\Omega \eta \mathcal{L} v_2 \, dx = \int_S \eta \mathcal{L} v_2 \, dx = \int_S \mathcal{L}\eta v_2 \, dx = (\mu, v_2).$$

By adding these equalities one obtains (44).

It remains to verify that (44) has a unique solution. Indeed the difference w of two solutions satisfies $w \in L^{p'}(\Omega)$ and

$$\int_\Omega w \mathcal{L} v \, dx = 0 \quad \forall v \in Y$$

where $Y = W^{2,p}(\Omega) \cap W_0^{1,p}(\Omega)$.

As $\mathcal{L}$ is an isomorphism from Y onto $L^p(\Omega)$, its adjoint $\mathcal{L}^*$ is an isomorphism from $L^{p'}(\Omega)$ onto Y'. One has

$$(\mathcal{L}^* w, v)_{Y'\times Y} = 0 \quad \forall v \in Y;$$

therefore $\mathcal{L}^* w = 0$ and $w = 0$.

v) *Necessary optimality condition.* From the definition of (q,h) one has

$$A^*(q,h)\cdot(\psi',\lambda') = (\psi_0 \delta_\Sigma, 0)\cdot(\psi',\lambda') = \int_\Sigma \psi_0 \psi' \, d\sigma.$$

Moreover, from the definition (30) of (ψ',λ') one has

$$(q,h)\cdot A(\psi',\lambda') = (q,h)\left(\frac{1_{C_1}}{|C_1|}, 1\right) = \frac{1}{|C_1|}\int_{C_1} q \, dx + h.$$

Then the optimality condition (42) shows that

$$\frac{1}{|C_1|}\int_{C_1} q \, dx = -h = q_\Gamma.$$

One can of course replace C_1 by any coil C_i, whence the announced result.

Remark 12. A sequential quadratic method is used in [3] in order to solve numerically this optimal control problem and it is applied to the control of the plasma shape in the Tokamaks JET and TORE SUPRA.

Acknowledgment. The authors are very thankful to Dr. H. Berestycki for many valuable comments on this paper.

REFERENCES

[1] H. BERESTYCKI AND H. BREZIS, *On a free boundary problem arising in plasma physics*, Nonlinear Anal., 4 (1980), pp. 415–436.

[2] J. BLUM, J. LE FOLL AND B. THOORIS, *The self consistent equilibrium and diffusion code* SCED, Comput. Phys. Comm., 24 (1981), pp. 235–254.

[3] _____, *Le contrôle de la frontière libre du plasma dans un Tokamak*, Proc. of the 5th International Conference on Analysis and Optimization of Systems. Versailles, 1982, Lecture Notes in Control and Information Sciences, 44, Springer, Berlin, pp. 852–867.

[4] L. A. CAFFARELLI AND A. FRIEDMAN, *Asymptotic estimates for the plasma problem*, Duke Math. J., 47 (1980), pp. 705–742.

[5] A. DERVIEUX, *Perturbation des équations d'équilibre d'un plasma confiné*, Rapport de recherche no. 18. INRIA, 1980.

[6] T. GALLOUËT, *Contribution à l'étude d'une équation apparaissant en physique des plasmas*, Thèse 3ème cycle, Université Paris VI, 1978.

[7] H. GRAD AND H. RUBIN, *Hydromagnetic equilibria and force-free fields*, Proc. Second International Conference on the Peaceful Uses of Atomic Energy, Geneva, Vol. 31, 190, United Nations, New York, 1958.

[8] H. GRAD, *Plasma transport in three dimensions*, Ann. New York Acad. Sci., 357 (1980), pp. 223–235.

[9] J. LERAY AND J. SCHAUDER, *Topologie et équations fonctionnelles*, Ann. Sci. Ecole Normale Sup., 51 (1934), pp. 45–78.

[10] C. MERCIER, *Lectures in plasma physics, The MHD approach to the problem of plasma confinement in closed magnetic configurations*, C.E.C. Luxembourg, 1974.

[11] J. P. PUEL, *Sur un problème de valeur propre non linéaire et de frontière libre*, CRAS Paris A, 284 (1977), pp. 861–863.

[12] D. G. SCHAEFFER, *Non-uniqueness in the equilibrium shape of a confined plasma*, Comm. PDE, 2 (6) (1977), pp. 587–600.

[13] V. D. SHAFRANOV, *On magnetohydrodynamical equilibrium configurations*, Soviet Physics JETP, 6 (33)(1958), pp. 545–554.

[14] V. D. SHAFRANOV AND V. S. MUKHOVATOV, *Plasma equilibrium in a Tokamak*, Nuclear Fusion, 11 (1971), pp. 605–633.

[15] J. SIMON, *Remarks on the plasma equilibrium problem with a limiter*, to appear.

[16] G. STAMPACCHIA, *Equations elliptiques du second ordre à coefficients discontinus*, Presses de l'Université de Montrèal, 1965.

[17] R. TEMAM, *A nonlinear eigenvalue problem: the shape at equilibrium of a confined plasma*, Arch. Rat. Mech. Anal., 60 (1975), pp. 51–73.

[18] _____, *Remarks on a free boundary value problem arising in plasma physics*, Comm. PDE 2 (6) (1977), pp. 563–585.

SIAM J. MATH. ANAL.
Vol. 17, No. 5, September 1986

009

THE SCALAR RIEMANN PROBLEM IN TWO SPATIAL DIMENSIONS: PIECEWISE SMOOTHNESS OF SOLUTIONS AND ITS BREAKDOWN*

W. B. LINDQUIST†

Abstract. Consider the scalar quasilinear equation

$$\frac{\partial u(t,x)}{\partial t}+\sum_{i=1}^{n}\frac{\partial f_i(u(t,x))}{\partial x_i}=0,$$

for $n=1,2$, $f_i\in C^2$: $R\to R$. For $n=2$, we define the two-dimensional Riemann problem and show the unique (in the sense of Kružkov) solutions are piecewise smooth for $f_1\equiv f_2\equiv f$, f purely convex or having a single inflection point. A mechanism leading to a presumed loss of piecewise smoothness is presented for f having three or more inflection points. The analysis is based on a study of the generalization of the one-dimensional Riemann problem to allow for initial data having a finite number of jump discontinuities with constant data or rarefaction waves between jumps.

Key words. Riemann problems, hyperbolic equations

AMS(MOS) subject classifications. Primary 35C05; secondary, 35B65, 35L65

1. Introduction.

Consider the Cauchy problem for the scalar quasilinear equation

$$\frac{\partial u(t,x)}{\partial t}+\sum_{i=1}^{n}\frac{\partial f_i(u(t,x))}{\partial x_i}=0,\qquad n=1,2,$$
$$u(0,x)=u_0(x),\tag{1.1}$$

with f_i: $\mathbf{R}\to\mathbf{R}\in C^2$. For two spatial dimensions, we define the Riemann problem and show that the unique (in the sense of Kružkov) solutions are piecewise smooth for $f_1\equiv f_2\equiv f$, f having at most one inflection point. The assumption $f_1=f_2$ appears to be natural from the point of view of applications. This sufficiency condition follows from a detailed analysis of a generalization of the Riemann problem in one dimension to allow for initial data having a finite number of jump discontinuities where the data between points of jump is allowed to contain rarefaction waves as defined below. For these generalized one-dimensional Riemann problems, we show the unique (in the sense of Kružkov) solutions are piecewise smooth if the initial data has a single jump discontinuity and f_1 is restricted to at most two inflection points. If the initial data has more than a single jump discontinuity we show that restricting f_1 to at most one inflection point guarantees a piecewise smooth solution.

We present a construction based on a function f_1 having three inflection points which shows a possible breakdown of piecewise smoothness for the associated one-dimensional Riemann problem. The implication of this example is that the scalar Riemann problem in two dimensions can presumably fail to have piecewise smooth solutions for nonconvex $f_2\equiv f_1$ having three inflection points.

*Received by the editors September 24, 1984, and in revised form March 1, 1985.

†Courant Institute of Mathematical Sciences, New York University, New York, New York, 10012. This research was supported in part by the Applied Mathematical Sciences subprogram of the Office of Energy Research, U. S. Department of Energy under contract DE-AC02-76ER03077.

A bounded measurable function $u(t,x)$ is a *weak solution* of (1.1) in the strip $\Pi_T \equiv [0,T]\times \mathbf{R}^n$ if

$$\iint_{\Pi_T} dt\,dx\left[u\frac{\partial\phi}{dt}+\sum_{i=1}^{2} f_i(u)\frac{\partial\phi}{\partial x_i}\right]+\int_{i=0} dx \quad u_0(x)\phi(0,x)=0, \tag{1.2}$$

for all $\phi\in C_0^\infty(\Pi_T)$. A function is a weak solution in the large if it is a weak solution for all T. For initial data $u_0(x)$ in the class of bounded, measurable functions Kružkov [2] has shown the existence and uniqueness of a bounded measurable function which is a weak solution to (1.1).[1] The uniqueness (entropy) condition satisfied by this weak solution is [2]

$$\iint_{\Pi_T} dt\,dx\,\operatorname{sign}(u-k)\left\{(u-k)\frac{\partial\phi}{\partial t}+\sum_{i=1}^{2}\left(f_i(u)-f_i(k)\right)\frac{\partial\phi}{\partial x_i}\right\}\geqq 0, \tag{1.3}$$

for any real constant k and any $\phi\in C_0^\infty(\Pi_T)$ such that $\phi\geqq 0$. This uniqueness condition characterizes the allowed discontinuities in the weak solution.

Define a *regular point* of the weak solution u as a point p for which $\partial u/\partial t$ and $\partial u/\partial x_i$ exist and are continuous in a neighbourhood of p, and at which the differential form of (1.1) is satisfied. The function u shall be termed *smooth* in an open neighbourhood N if each point $p\in N$ is a regular point. Let a *smooth* surface of codimension i in $\mathbf{R}^n\times\mathbf{R}^+$ be a surface at each point p of which a unique (modulo minus signs) i-dimensional space of normals exist and for which the normal space varies continuously as the point p is varied over the surface. A *piecewise smooth surface* is defined by induction on the codimension i, as the union of a finite number of smooth surfaces with piecewise smooth boundaries of codimension $i+1$; each of the finite number of pieces being everywhere the boundary of the same codimension i surfaces. A *piecewise smooth* solution $u\in\mathbf{R}^n\times\mathbf{R}^+$ is defined as a function that is smooth except on a discontinuity set that consists of piecewise smooth surfaces. A *point of jump* is a point p of discontinuity of a function u for which the discontinuity is a simple jump in u, i.e. locally the point p lies on a smooth codimension 1 surface of discontinuity and the one sided limits $u^-=\lim_{\varepsilon\to 0^+} u(p-\varepsilon n)$ and $u^+=\lim_{\varepsilon\to 0^+} u(p+\varepsilon n)$, where n is the unit normal to the smooth discontinuity surface at p, exist at p. A singular point of u which is not a point of jump shall be termed an *irregular* point.

Using integration by parts, the arbitrariness of ϕ and appropriate choices of the constant k, it can be shown that (1.3) is equivalent to two conditions for points of jump p in whose neighbourhood the unique weak solution u is piecewise smooth:

$$n\cdot\left[u^+-u^-,\vec{f}(u^+)-\vec{f}(u^-)\right]=0, \tag{1.4}$$

and

$$n\cdot\left[k-u^-,\vec{f}(k)-\vec{f}(u^-)\right]\geqq 0, \tag{1.5}$$

where $\vec{f}$ stands for the vector $[f_1,f_2]$. In (1.5) n is oriented such that $u^-\leqq u^+$ and k is any constant such that $u^-\leqq k\leqq u^+$. As a consequence of (1.4), (1.5) also holds with u^- replaced by u^+. Equation (1.4) will be denoted the jump condition for the discontinuity at p and (1.5) the entropy condition. In $\mathbf{R}^1\times\mathbf{R}^+$ (1.4) is the familiar Rankine–Hugoniot

[1] Kružkov's proof holds for $n\geqq 1$ spatial dimensions. Here only $n\leqq 2$ concerns us.

condition and (1.5) is equivalent to the entropy condition of Oleĭnik [5]. The advantage of equations (1.4) and (1.5) over (1.3) is that the former are local conditions which will allow construction of piecewise smooth global solutions by piecing together local solutions which individually obey (1.4) and (1.5).

A weak solution $u \in \mathbf{R}^n \times \mathbf{R}^+$ to (1.1) which is piecewise smooth and whose discontinuity sets are piecewise smooth surfaces of points of jump obeying (1.4) and (1.5) then satisfies (1.3) and is therefore the unique Kružkov solution to (1.1).

The existence and uniqueness theory gives little insight into the actual form of the weak solution. In $\mathbf{R}^1 \times \mathbf{R}^+$, insight is provided by analysis of the Riemann problem whose solution is piecewise smooth and can be derived in terms of a nonlinear wave analysis. It can be shown (see, for example, Lax [3] and the references therein) that these nonlinear waves consist of rarefaction and shock waves. The initial motivation of this paper was to consider the natural generalization of the Riemann problem in two dimensions and investigate its solution in terms of two-dimensional waves which, by implication, is a restriction to the set of piecewise smooth solutions. In this paper we determine a sufficient condition for which the solution to the two-dimensional Riemann problem is piecewise smooth. In a companion paper [4] we investigate the form of these solutions in terms of nonlinear waves. For the case $f_1 = f_2 \equiv f$, the sufficiency condition consists of restricting f to at most one inflection point. We conjecture that if the number of inflection points of f is at least three, the solution to the generalized Riemann problem defined in one dimension (and hence for the Riemann problem in two dimensions) may fail to be piecewise smooth. Our argument for this is presented in §2.

The generalizations of the one-dimensional Riemann problem to allow multiple jumps and restricted smooth variation in the initial data are presented in §§2 and 3. In §4, the two-dimensional Riemann problem is defined and a sufficiency condition for piecewise smooth solutions obtained.

2. The one-dimensional Riemann problem. Throughout this paper the notation $q'(p)$ or $(q(p))'$ shall denote $dq(p)/dp$ for any function q of a single variable p.

We first discuss the familiar results for the one-dimensional Riemann problem to introduce notation and reformulate results in a manner that will allow generalization. The scalar quasilinear equation (1.1) in one spatial dimension ($f_1 \equiv f, x_1 \equiv x$) with initial data

$$u(0,x) = \begin{cases} u_l & \text{for } x < x_0, \\ u_r & \text{for } x > x_0, \end{cases} \tag{2.1}$$

defines the *one-dimensional Riemann problem*.

The Riemann problem (1.1), (2.1) is invariant under the similarity transformation $(t, x - x_0) \to (ct, c(x - x_0))$, $c > 0$.[2] Hence the solution is constant along the straight lines $(x - x_0)/t = constant$. In a neighbourhood of a point in which the Riemann problem solution is smooth, (1.1) implies that the solution will be constant along the straight lines

$$\frac{x - x_c}{t - t_c} = f'(u), \qquad t \geq t_c, \tag{2.2}$$

[2] This scale invariance is one of the essential features of the Riemann problem.

which are denoted *characteristics*. Characteristics shall be denoted (u,t_c,x_c) where (t_c,x_c) is the origin point of the characteristic given by (2.2). For the Riemann problem (1.1) (2.1), $t_c=0$ for all characteristics.

Let $p=(t_p,x_p)$ be a point on a characteristic (u_1,t_1,x_1). Let $\hat{n}$ be a normal to the characteristic at p. By the smoothness of u there exists δ_p such that each point $p+\varepsilon\hat{n}$, where $-\delta_p\leqq\varepsilon\leqq\delta_p$, lies on a characteristic. Let $(u_\varepsilon,t_\varepsilon,x_\varepsilon)$ denote such a characteristic. Then (u_1,t_1,x_1) is a *rarefaction wave* if the lines defined by the characteristics (u_1,t_1,x_1) and $(u_\varepsilon,t_\varepsilon,x_\varepsilon)$ intersect at some $t\leqq t_p$. This condition generalizes the concept of the familiar centered rarefaction wave for which $t_\varepsilon=t_1$, $x_\varepsilon=x_1$, to noncentered rarefaction waves which appear in the generalized Riemann problem solutions and also generalizes to a definition of rarefaction waves in two spatial dimensions.

A *rarefaction fan* is an open set in the t,x plane all points of which belong to rarefaction waves.

A *shock* is defined as a smooth curve of points of jump in the t,x plane at each point of which the conditions (1.4) and (1.5) are satisfied. A shock curve (wave) will be denoted $(u_l(t),u_r(t),t_s,x_s)$ where $u_l(t)$ is the limit value of the solution u to the left of the shock wave at time $t>t_s$, u_r is the corresponding right limit, and (t_s,x_s) is the origin point of the shock wave. For the Riemann problem (1.1) (2.1), $t_s=0$, $x_s=x_0$.

A *constant state* is a domain (connected open set) in the t,x plane over which the solution is constant.

Let $f(u)\in C^2$: $\mathbf{R}\to\mathbf{R}$. Let $\{u_1,u_2\}$ denote a closed interval where

$$\{u_1,u_2\}\equiv\begin{cases}[u_1,u_2] & \text{if } u_1<u_2,\\ [u_2,u_1] & \text{if } u_1>u_2.\end{cases} \tag{2.3}$$

A *rarefaction interval of* f shall denote an interval which satisfies either

$$f''(u)>0 \quad \text{if } u_1<u_2 \text{ for every } u\in\{u_1,u_2\}, \tag{2.4a}$$

or

$$f''(u)<0 \quad \text{if } u_1>u_2 \text{ for every } u\in\{u_1,u_2\}. \tag{2.4b}$$

Let $E(u_1,u_2;u)$: $\mathbf{R}\to\mathbf{R}$ denote the convex envelope function for $f(u)$ restricted to the interval $\{u_1,u_2\}$ of f such that

$$E(u_1,u_2;u)\equiv\begin{cases}\text{upper convex envelope from } u_1 \text{ to } u_2, & \text{if } u_1>u_2,\\ \text{lower convex envelope from } u_1 \text{ to } u_2 & \text{if } u_1<u_2.\end{cases} \tag{2.5}$$

Let $C(u_1,u_2;u)$: $\mathbf{R}\to\mathbf{R}$ denote the linear map restricted to the domain interval $\{u_1,u_2\}$ such that

$$\begin{aligned}C(u_1,u_2;u_1)&=f(u_1),\\ C(u_1,u_2;u_2)&=f(u_2),\end{aligned} \tag{2.6}$$

i.e. for all u in $\{u_1,u_2\}$,

$$\frac{dC(u_1,u_2;u)}{du}=\frac{f(u_2)-f(u_1)}{u_2-u_1}.$$

Denote the restriction of the function $f(u)$ to the subinterval $\{u_1,u_2\}$ by $f(u_1,u_2;u)$.

THEOREM 2.1 (Oleĭnik [5]). *Consider the Riemann problem* (1.1), (2.1). *Assume* $f(u)\in C^2$: $\mathbf{R}\to\mathbf{R}$ *such that f has a finite number of inflection points. Then the unique, entropy obeying solution connecting the states* u_l *and* u_r *is determined solely from* $E(u_l,u_r;u)$, *as described below. Further, the solution is piecewise smooth, with a single irregular point corresponding to the discontinuity at* x_0 *in the initial data.*

The convex envelope function $E(u_l,u_r;u)$ uniquely divides the interval $\{u_l,u_r\}$ into rarefaction intervals separated by intervals corresponding to the linear chord segments of E. Each rarefaction interval corresponds to a rarefaction fan in the solution. The fans are centered on $t=0$, $x=x_0$ and each rarefaction wave has a unique value u corresponding to a state in a rarefaction interval. The slope of the rarefaction wave $(v,0,x_0)$ is given by $f'(v)$. Each linear chord in E covering an interval $\{u_1,u_2\}$ corresponds to a shock wave $(u_1,u_2,0,x_0)$. The solution is bordered on the left by the single constant state $u=u_l$ and on the right by $u=u_r$. No constant states appear within the solution except for these two bounding constant states, which are the states u_l and u_r of the initial data. The speeds of the nonlinear waves found in the solution increase continuously from left to right along the x direction. Fig. 2.1a shows an upper convex envelope drawn between two states for the function $f(u)$ illustrated. Figure 2.1b illustrates the corresponding Riemann solution in t,x space, with the three shocks (in this case) represented as dark lines.

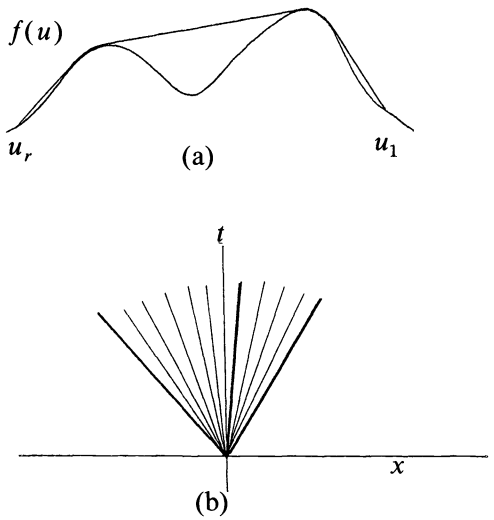


FIG. 2.1. (a) *The upper convex envelope drawn between two points* u_l *and* u_r *for an illustrative function f.* (b) *The Riemann solution in the* t,x *plane based on* (a). *Dark lines represent shock waves, light lines are rarefaction waves.*

A solution given by Theorem 2.1 shall be referred to, for brevity, as the *Oleĭnik solution to* $\{u_l,u_r\}$ centered on x_0.

2.1. The generalized Riemann problem for initial data with a single point of discontinuity. Let $h(x_1,x_2,u_1,u_2;x)$: $\mathbf{R}\to\mathbf{R}$ be a continuous, differentiable, 1-1 mapping from the interval $[x_1,x_2]$ of the real axis onto a rarefaction interval $\{u_1,u_2\}$ of f such

that

$$h(x_1,x_2,u_1,u_2;x_1)=u_1, \qquad h(x_1,x_2,u_1,u_2;x_2)=u_2. \tag{2.7}$$

Further, assume that the derivative of h is bounded and bounded away from 0. Then its inverse map exists, is 1-1, onto, and has derivatives bounded away from 0. Denote the inverse map by

$$h^{-1}(x_1,x_2,u_1,u_2;u).$$

Where no ambiguity exists the notation for these maps shall be shortened to $h(x)$ and $h^{-1}(u)$.

Let $f(u)\in C^2$: $\mathbf{R}\to\mathbf{R}$ such that $f''(u)=0$ at a finite number of points. Let $\{a,b\}$ and $\{c,d\}$ be rarefaction intervals of $f(u)$.

DEFINITION 2.2. The Cauchy problem

$$\frac{\partial u(x,t)}{\partial t}+\frac{\partial f(u(x,t))}{\partial x}=0, \tag{2.8}$$

with initial data

$$u(x)=\begin{cases} a, & x\leqq x_l,\\ h_l(x_l,x_0,a,b;x), & x_l<x<x_0,\\ h_r(x_0,x_r,c,d;x), & x_0<x<x_r,\\ d, & x_r\leqq x, \end{cases} \tag{2.9}$$

will be denoted as a Cauchy problem with generalized Riemann data (generalized Riemann problem, for short). The unique solution to (2.8) and (2.9) obeying (1.4) and (1.5) shall be referred to as the generalized Oleĭnik solution.

THEOREM 2.3. *Let $f(u)\in C^2$: $\mathbf{R}\to\mathbf{R}$ such that $f''(0)$ for at most two values of u. Then the generalized Oleĭnik solution to* (2.8), (2.9) *is piecewise smooth; it consists of two constant states, rarefaction fans, at most three shock waves*[3] *and at most two irregular points. A rarefaction wave either originates at $t=0$ on the interval $[x_l,x_r]$ or tangentially to a shock line at $t>0$ and either terminates on a shock line or propagates undisturbed for all t. At most two points can be irregular, the initial discontinuity point in the data $(t=0,x=x_0)$ or the meeting point of three shocks.*[3]

2.1.1. Proof of Theorem 2.3. The following definitions will prove useful. Let $\{u_1,u_2\}$ be a rarefaction interval of f and $h(x_1,x_2,u_1,u_2;x)$ a map as in (2.7). At each point $x\in[x_1,x_2]$ construct the rarefaction wave $(u,0,x)$ with

$$x(t)=x+f'(h(x))t.$$

The set of waves thus constructed on $[x_1,x_2]$ form a rarefaction fan which shall be referred to as the *rarefaction fan determined by* $h(x_1,x_2,u_1,u_2;x)$.

[3] By definition of a shock wave as a smooth curve, the occurrences of either two shock lines interacting to produce a single shock curve, or a single shock curve splitting into two shock curves, are each described as involving three shock waves.

Let $\{u_1, u_2\}$ be a rarefaction interval of f. For every u in $\{u_1, u_2\}$ construct the rarefaction wave $(u, 0, x_0)$ with

$$x(t) = x_0 + f'(u)t.$$

The set of waves thus constructed form a rarefaction fan which shall be referred to as *determined at x_0 by the interval $\{u_1, u_2\}$*.

Let $\{a, b\}$ be a rarefaction interval of f. Let $h(\rho_0, \rho_1, a, b; \rho)$ be a map as in (2.7) from the real interval $[\rho_0, \rho_1]$ to $\{a, b\}$. Construct a curve segment $(t(\rho), x(\rho))$ from (t_0, x_0) to (t_1, x_1) in the t, x plane such that

$$\begin{aligned} (t_0, x_0) &\equiv (t(\rho_0), x(\rho_0)), \\ (t_1, x_1) &\equiv (t(\rho_1), x(\rho_1)), \\ \frac{dx(\rho)}{dt(\rho)} &= f'(h(\rho_0, \rho_1, a, b; \rho)). \end{aligned}$$

At each point $(t(\rho), x(\rho))$ construct the rarefaction wave $(h(\rho), t(\rho), x(\rho))$. Then the set of waves constructed on the curve segment form a rarefaction fan which shall be referred to as the *rarefaction fan determined by the curve $(t(\rho), x(\rho))$*.

LEMMA 2.4. *Let N be a rarefaction fan in the t, x plane for some solution $u(t, x)$. Let $[x_1, x_2]$ be an interval in N parallel to the x-axis at some time t such that the rarefaction wave passing through the point (t, x_1) has value $u = u_1$ and the rarefaction wave passing through the point (t, x_2) has value $u = u_2$. Then $\{u_1, u_2\}$ is a rarefaction interval of f and the fan defines a continuous*, 1-1 *mapping $h'(x_1, x_2, u_1, u_2; x)$: $\mathbf{R} \to \mathbf{R}$ from $[x_1, x_2]$ onto $\{u_1, u_2\}$ whose derivatives are bounded and bounded away from* 0.

The proof for Lemma 2.4 follows from the definition of a rarefaction wave. □

The proof for Theorem 2.3 is constructive. All possible Cauchy data (2.9) fall into four classes:

C1. $\{a, b\}$ and $\{c, d\}$ are not disjoint, $f'' > 0$ on $\{a, b\} \cup \{c, d\}$,
C2. $\{a, b\}$ and $\{c, d\}$ are not disjoint, $f'' < 0$ on $\{a, b\} \cup \{c, d\}$,
C3. $\{a, b\}$ and $\{c, d\}$ are disjoint, $b > c$,
C4. $\{a, b\}$ and $\{c, d\}$ are disjoint, $b < c$.

Classes C3 and C4 can each be divided into four subgroups:

S1: $a < b$, $c < d$,
S2: $a > b$, $c > d$,
S3: $a < b$, $c > d$,
S4: $a > b$, $c < d$.

For classes C1, C2 there are no inflection points of f in the interval $[a, d]$. For subgroups S1 and S2 of classes C3 and C4 there are either 0 or 2 inflection points in the interval $[a, d]$. Subgroups S3 and S4 of classes C3 and C4 contain one inflection point in the interval $[a, d]$. Theorem 2.3 will be proven for the case of 0 and 1 inflection point in the interval $[a, d]$ of f. The proof for the case of two inflection points will be omitted for brevity.

f has zero inflection points. Restriction of f to no inflection points implies initial data from either classes C1, C2 or subgroups S1 or S2 of C3 and C4. We study data of classes C1, C3.S1 and C3.S2; the proofs for classes C2, C4.S1 and C4.S2 follow analogously.

Class C1. There are four arrangements of initial data in the class C1:

$$\begin{aligned} &\text{C1.A1:} \quad a<c<b<d,\\ &\text{C1.A2:} \quad a<c<d<b,\\ &\text{C1.A3:} \quad c<a<d<b,\\ &\text{C1.A4:} \quad c<a<b<d. \end{aligned}$$

The generalized Oleĭnik solution to Cauchy data of type C1.A1 is examined below.

Let $\{a,b\}$ and $\{c,d\}$ be rarefaction intervals of type C1.A1. Consider maps $h_l(x_l,x_0,a;b;x)$ and $h_r(x_0,x_r,c,d;x)$ as in (2.7). Construct the rarefaction fans determined by $h_l(x)$ and $h_r(x)$. Then through each point $p\equiv(t,x)$ in the plane such that

$$x_0+f'(c)t\leqq x\leqq x_0+f'(b)t$$

there passes a unique rarefaction wave originating from the interval $[x_l,x_0]$ and a unique rarefaction wave originating from the interval $[x_0,x_r]$, thus defining respectively the unique values $u_l(p)$ and $u_r(p)$.

Consider the curve $(t,x(t))$ defined implicitly by

$$x(t)=x_0+\int_0^t \frac{dx}{dt}\,dt, \tag{2.10}$$

where

$$\frac{dx}{dt}=\frac{f(u_l(t))-f(u_r(t))}{u_l(t)-u_r(t)}, \tag{2.11}$$

with $u_l(t)$ and $u_r(t)$ determined from

$$x(t)=h_l^{-1}(x_l,x_0,a,b;u_l(t))+f'(u_l(t))t, \tag{2.12}$$

$$x(t)=h_r^{-1}(x_0,x_r,c,d;u_r(t))+f'(u_r(t))t. \tag{2.13}$$

We remark that for $\{a,b\}$ and $\{c,d\}$ of type C1.A1, the set of equations (2.10) through (2.13) are well defined for all t and the curve $(t,x(t))$ is smooth.

PROPOSITION 2.5. *For the curve* $(t,x(t))$ *defined by* (2.10) *through* (2.13):

1) $du_l(t)/dt<0$, $du_r(t)/dt>0$, $\forall t>0$.

2) *There exists* $g\in[c,b]$ *such that* $u_l(t)\to g$, $u_r(t)\to g$ *as* $t\to\infty$.

3) *The curve defined in* (2.10) *through* (2.13) *considered as a curve of points of jump* $u_l(t)\to u_r(t)$ *obeys* (1.4) *and* (1.5).

Proof. Note

$$\begin{aligned} &f''(u_l(t))>0, \qquad \big(h_l^{-1}(u)\big)'>0 \quad \forall u_l(t) \text{ on } [a,b],\\ &f''(u_r(t))>0, \qquad \big(h_r^{-1}(u)\big)'>0 \quad \forall u_r(t) \text{ on } [c,d]. \end{aligned}$$

Equations (2.12) and (2.13) yield

$$\big[f'(u_l(t))-f'(u_r(t))\big]t=h_r^{-1}(u_r(t))-h_l^{-1}(u_l(t)). \tag{2.14}$$

For $t>0$, the positivity of the right-hand side of (2.14) implies

$$f'(u_l(t))-f'(u_r(t))>0. \tag{2.15}$$

By the requirement that $\{a,b\}$ and $\{c,d\}$ be of class C1.A1, (2.15) implies (using 2.11)

$$f'(u_l(t))>\frac{dx}{dt}>f'(u_r(t)). \tag{2.16}$$

Therefore, the time derivatives of (2.12) and (2.13),

$$\frac{dx}{dt} = \left[\left(h_l^{-1}(u_l)\right)' + f''(u_l)t\right]u_l'(t) + f'(u_l), \tag{2.17}$$

$$\frac{dx}{dt} = \left[\left(h_r^{-1}(u_r)\right)' + f''(u_r)t\right]u_r'(t) + f'(u_r), \tag{2.18}$$

lead to the conclusion

$$u_l'(t) < 0, \qquad u_r'(t) > 0 \quad \forall t. \tag{2.19}$$

Further (2.16) and (2.19) imply the existence of $g \in [c,b]$ such that as $t \to \infty$,

$$u_l(t) \to g, \quad u_r(t) \to g, \quad \frac{dx}{dt} \to f'(g).$$

By construction, the curve $(t,x(t))$ considered as a curve of points of jump obeys (1.4). It is easy to verify that the entropy condition (1.5) is also satisfied. From Proposition 2.5 note that the strength of the shock decreases monotonically to zero as $t \to \infty$. □

All rarefaction waves originating in the interval

$$h_l^{-1}(g) \leqq x \leqq h_r^{-1}(g)$$

for the construction in Proposition 2.5 intersect the curve $(t,x(t))$. Terminate these waves at the curve. All rarefaction waves originating in the intervals

$$h_l^{-1}(a) \leqq x \leqq h_l^{-1}(g),$$
$$h_r^{-1}(g) \leqq x \leqq h_r^{-1}(d),$$

propagate unimpeded for all $t > 0$. The rarefaction fans and shock curve thus defined by this construction shall be said to be *determined by the functions* $h_l(x_l, x_0, a, b; x)$ *and* $h_r(x_0, x_r, c, d; x)$.

PROPOSITION 2.6. *Let* $\{a,b\}$ *and* $\{c,d\}$ *be as in* C1.A1. *Then the generalized Oleĭnik solution* $u(t,x)$ *to* (2.8) (2.9) *is given by*:

the constant state $u = a$ *for* $x \leqq x_l + f'(a)t$, $t \geqq 0$;

the shock curve and rarefaction fans determined by $h_l(x)$ *and* $h_r(x)$ *for* $x_l + f'(a)t < x < x_r + f'(d)t$, $t \geqq 0$;

the constant state $u = d$ *for* $x_r + f'(d)t \leqq x$, $t \geqq 0$.

The proof follows immediately from Proposition 2.5. □

A sketch of this solution is given in Fig. 2.2.

The generalized Oleĭnik solutions to Cauchy data of types C1.A2 through C1.A4 have exactly the same structure as for C1.A1. For each, the strength of the single shock curve $(t,x(t))$ decreases monotonically to zero as $t \to \infty$, and $dx/dt \to f'(g)$ where for

C1.A2: $c < g \leqq d$,
C1.A3: $a < g \leqq d$,
C1.A4: $a < g \leqq b$.

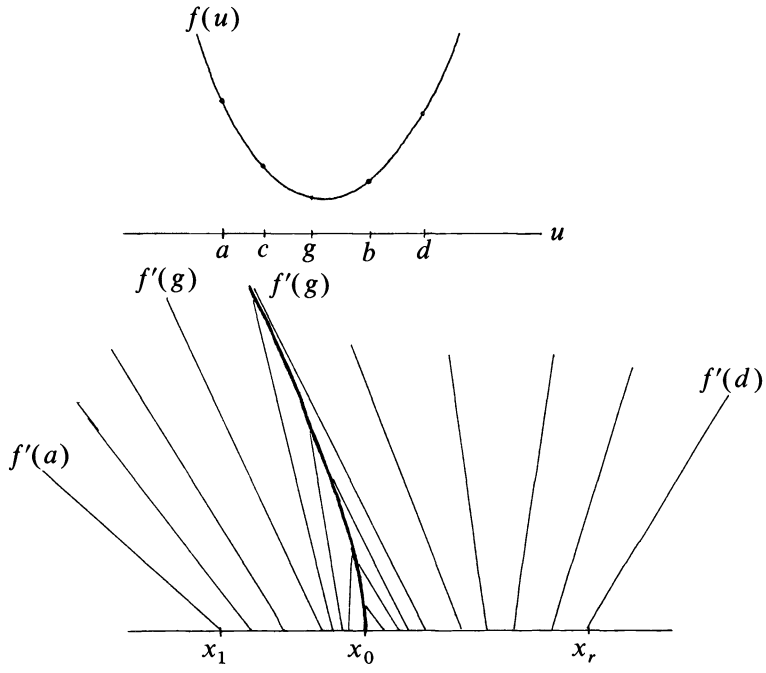


FIG. 2.2. *A sketch of the solution (lower figure) for a particular case of initial data (upper figure) for a generalized Riemann problem with f having no inflection point. The dark line is the shock wave. Important rarefaction lines have been labelled by their speeds.*

Class C3.S1: no inflection point. Let $\{a,b\}$ and $\{c,d\}$ be rarefaction intervals of type C3.S1 such that for every $v\in\{a,b\}$, $w\in\{c,d\}$.

$$E(v,w;u)=C(v,w;u).$$

Then $f'(v)>f'(w)$ for every choice v,w. Consider the curve $(t,x(t))$ implicitly defined by (2.10) through (2.13). Statement 1 of Proposition 2.5 follows from (2.14) through (2.19). However, the disjointness of the intervals implies there exists finite times $t_l>0$, $t_r>0$ such that $u_l(t_l)=a$ and $u_r(t_r)=d$. For all $t>t_l$ let $u_l(t)\equiv a$ and for all $t>t_r$ let $u_r(t)\equiv d$. With this extension, (2.10) through (2.13) are defined for all t. The curve $(t,x(t))$ thus defined, considered as a curve of points of jump $u_l(t)\to u_r(t)$ obeys the conditions (1.4) and (1.5). In contrast to class C1 data, the shock strength attains a constant value, $|a-d|$, as $t\to\infty$.

All rarefaction waves originating from the interval $[x_l,x_r]$ intersect the shock curve and are terminated there. The shock curve and rarefaction fans so defined (together with the bounding constant states $u=a$, $u=d$ will also be referred to as *determined by the functions* $h_l(x)$ *and* $h_r(x)$ since their definition is merely an extension of the usage defined after Proposition 2.5.

PROPOSITION 2.7. *Let* $\{a,b\}$ *and* $\{c,d\}$ *be disjoint rarefaction intervals on f such that for all* $v\in\{a,b\}$, $w\in\{c,d\}$

$$E(v,w;u)=C(v,w;u).$$

Then the generalized Oleĭnik solution to (2.8) (2.9) *is given by constructing the shock curve and rarefaction fans determined by* $h_l(x_l,x_0,a,b;x)$ *and* $h_r(x_0,x_r,c,d;x)$.

Proof. The above construction proves this proposition for rarefaction intervals in class C3.S1, where the condition $f''>0$ was used implicitly in (2.17) and (2.18). With trivial modifications, analogous arguments can be used to prove this proposition for appropriate data in classes C3 and C4.

Class C3.S2: *no inflection point.* Data of class C3.S2 insures $f'(b)\leqq f'(c)$.

PROPOSITION 2.8. *Let* $\{a,b\}$ *and* $\{c,d\}$ *be rarefaction intervals of* f *of class* C3. *Further assume* $f'(b)\leqq f'(c)$. *Then the generalized Oleĭnik solution* $u(t,x)$ *to* (2.8) (2.9) *is*:

the constant state $u(t,x)=a$ *for* $x<x_l+f'(a)t$, $t\geqq 0$;

the rarefaction wave determined by $h_l(x_l,x_0,a,b;x)$ *for*

$$x_l+f'(a)t<x<x_0+f'(b)t, \qquad t\geqq 0;$$

the Oleĭnik solution to $\{b,c\}$ *centered on* x_0 *for* $x_0+f'(b)t<x<x_0+f'(c)t$, $t\geqq 0$;

the rarefaction wave determined by $h_r(x_0,x_r,c,d;x)$ *for*

$$x_0+f'(c)t<x<x_r+f'(d)t, \qquad t\geqq 0,$$

and the constant state $u(t,x)=d$ *for* $x<x_r+f'(d)t$, $t\geqq 0$.

Proof. The proof is trivial; the condition $f'(b)\leqq f'(c)$ guarantees that the rarefaction waves determined by h_l and h_r do not interact.

Thus all solutions for allowed Cauchy data with f restricted to have no inflection point are piecewise smooth, with at most a single shock and at most a single irregular point at $(t=0,x=x_0)$.

f has one inflection point. Restriction of f to a single inflection point implies initial data from subgroups S3 or S4 of C3 and C4. We study data of C3.S3 and C3.S4; the proofs for C4.S3 and C4.S4 follow analogously.

Class C3.S3.

LEMMA 2.9. *Let* f, $\{a,b\}$, *and* $\{b^*,a^*\}$ *be chosen such that*: $\{a,b\}$ *is a rarefaction interval of* f *on which* $f''(u)>0$, $\{b^*,a^*\}$ *is a rarefaction interval of* f *on which* $f''(u)<0$, $a^*<b$, *and*

$$E(b,b^*;u)=C(b,b^*;u), \quad f'(b^*)=\frac{dC(b,b^*;u)}{du},$$
$$E(a,a^*;u)=C(a,a^*;u), \quad f'(a^*)=\frac{dC(a,a^*;u)}{du}.$$

For every $v\in\{a,b\}$ *there exists* $w(v)\in\{b^*,a^*\}$ *such that*

$$E(v,w(v);u)=C(v,w(v);u), \qquad f'(w(v))=\frac{dC(v,w(v);u)}{du}.$$

Then

1) $w(v)$ *is a* 1-1 *map from* $\{a,b\}$ *onto* $\{b^*,a^*\}$,

2) $f''(w(v))dw(v)/dv=(f'(w(v))-f'(v))/(w(v)-v)$.

Proof. 1) This follows from the definition of $w(v)$.

2) For all $v\in(a,b)$, δv sufficiently small,

$$f(w(v))-f(v)=f'(w(v))[w(v)-v], \tag{2.20}$$
$$f(w(v+\delta v))-f(v+\delta v)=f'(w(v+\delta v))[w(v+\delta v)-v-\delta v]. \tag{2.21}$$

Subtracting (2.20) from (2.21), arranging terms and taking $\delta v \to 0$

$$f'(w(v))\frac{dw(v)}{dv} - f'(v) = f''(w(v))\frac{dw(v)}{dv}w(v) + f'(w(v))\frac{dw(v)}{dv} - f''(w(v))\frac{dw(v)}{dv}v - f'(w(v)),$$

from which the conclusion 2) follows. □

Let $\{a,b\}$ and $\{b^*,a^*\}$ be as in Lemma 2.9. Let $h_l(x_l,x_0,a,b;x)$ be the usual map from the rarefaction interval $\{a,b\}$ to $[x_l,x_0]$. Consider the curve $(t,x(t))$ defined by

$$x(t) = x_0 + \int_0^t \frac{dx}{dt}\,dt, \tag{2.10}$$

$$\frac{dx(t)}{dt} = f'(w(u_l(t))), \tag{2.22}$$

with $w(u_l(t))$ defined in Lemma 2.9 and $u_l(t)$ defined by

$$x(t) = h_l^{-1}(x_l,x_0,a,b;u_l(t)) + f'(u_l(t))t. \tag{2.12}$$

There exists a finite time $t_l > 0$ at which $u_l(t_l) = a$ and

$$\frac{dx}{dt} = f'(w(a)) = f(a^*).$$

For every $t > t_l$ let $u_l(t) = a$. Then (2.10), (2.12) and (2.22) are defined for all t.

PROPOSITION 2.10. *The curve defined by* (2.10), (2.12) *and* (2.22) *considered as a curve of points of jump* $u_l \to w(u_l)$ *obeys* (1.4) *and* (1.5).

Proof. Follows from the construction. □

Construct the rarefaction fan determined by h_l. If a wave from this fan intersects the above shock curve, terminate the wave on the curve. For a point $(t_b,x(t_b))$ on this shock curve where $0 \leqq t_b \leqq t_l$, there exists a uniquely defined $u_r(t_b) = w(u_l(t_b))$. Also note from (2.22) and Lemma 2.9

$$x''(t) = f''(w(u_l))w'(u_l)u_l'(t) = \frac{f'(w(u_l)) - f'(u_l)}{w(u_l) - u_l}u_l'(t) < 0. \tag{2.23}$$

The rarefaction fan determined by this curve can therefore be constructed. Note further that the "left bounding" wave $(a^*,t,x(t))$ of the fan propagates parallel to the shock curve.

The two rarefaction fans and shock wave, together with the bounding constant state $u = a$, thus constructed from (2.10), (2.12), (2.22) and (2.23) shall be denoted *as determined by the function* $h_l(x)$ *and the state* $u = b^*$.

Let f have a single inflection point and $\{a,b\}$ and $\{c,d\}$ be rarefaction intervals of type C3.S3. Then there exists a rarefaction interval $\{b^*,a^*\}$ as described in Lemma 2.9. Three possibilities exist for the orientation of $\{c,d\}$ with respect to $\{b^*,a^*\}$:

Case 1. $\{c,d\}$ and $\{b^*,a^*\}$ are disjoint with $d \geqq a^*$.

Case 2. $\{c,d\}$ and $\{b^*,a^*\}$ are disjoint with $c \leqq b^*$.

Case 3. $\{c,d\}$ and $\{b^*,a^*\}$ are not disjoint.

For the case $d \geqq a^*$, the generalized Oleĭnik solution $u(t,x)$ to (2.8) (2.9) is given by Proposition 2.7.

For the case $c \leqq b^*$, the generalized Oleĭnik solution $u(t,x)$ to (2.8) (2.9) is given by the following:

The constant state $u=a$, two rarefaction fans and the shock curve determined by the function $h_l(x)$ and the state $u=b^*$ for $x \leqq f'(b^*)t$, $t>0$.

The rarefaction wave determined at x_0 by the interval $\{c,b^*\}$ for $f'(b^*)t \leqq x \leqq f'(c)t$, $t>0$.

The rarefaction wave determined by $h_r(x)$ for $f'(c)t \leqq x \leqq f'(d)t$, $t>0$.

The constant state $u=d$ for $f'(c)t \leqq x$, $t>0$.

This solution is sketched in Fig. 2.3.

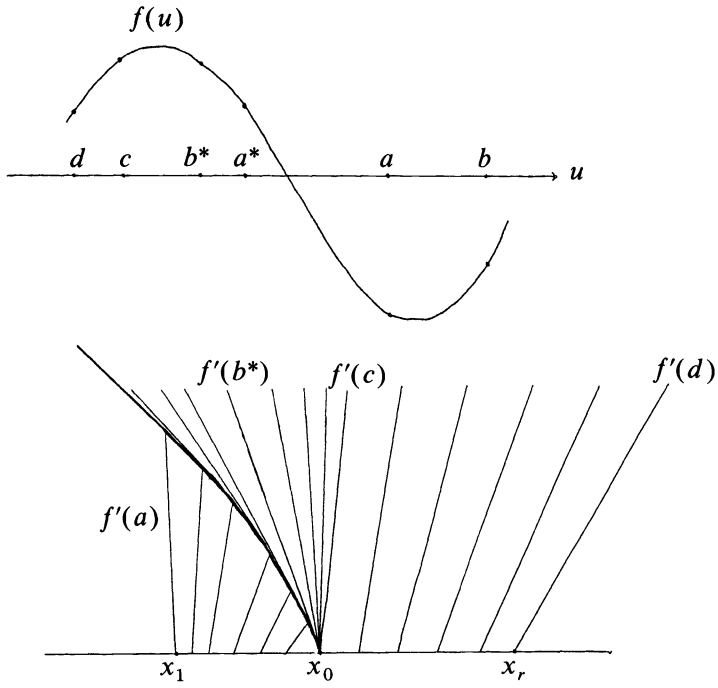


FIG. 2.3. *A sketch of the solution (lower figure) for a particular case of initial data (upper figure) for a generalized Riemann problem with f having a single inflection point. The dark line is the shock wave. Important rarefaction lines have been labelled by their speeds.*

For the case in which $\{c,d\}$ and $\{b^*,a^*\}$ are not disjoint we remark that the generalized Oleĭnik solution can be described by a solution similar to case 1 above for $0 \leqq t \leqq t_g$ and by a solution similar to case 2 for $t_g < t$. The time t_g is determined by the propagation of a rarefaction wave for some state $u=g$ where $g \in \{b^*,a^*\} \cap \{c,d\}$.

Class C3.S4. Lemma 2.9 and Proposition 2.10 can be written analogously for data of type C3.S4, with the arguments revolving around an interval $\{c^*,d^*\}$. Again the construction of a solution to the generalized Riemann problem can be broken into three cases and dealt with as above.

The solutions for classes C4.S3 and C4.S4 follow in the same manner. Thus we have shown that the solutions to the generalized Riemann problem for f having at most a single inflection point are piecewise smooth and consist of at most a single shock wave and at most a single irregular point (at x_0).

f has two inflection points. In the interest of brevity, we omit the solution for the case of f having two inflection points. The method of proof is repetitive to that presented above. The proof shows that in addition to solutions with a single shock, either two shock waves can be present for all $t>0$ or three shock waves can appear in the solution (i.e., a single shock which is present for $0 \leqq t \leqq t_s$ which then "splits", via interaction with rarefaction waves, into two shocks which are present for $t_s \leqq t$).

2.2. A conjecture on loss of piecewise smoothness for more than two inflection points. The constructive proof given for Theorem 2.3 is not readily extendable to f having three inflection points since the number of cases to be considered is large. In fact, we argue that for $f(u)$ having three or more inflection points, the solution need not be piecewise smooth. An example is sketched below in which a solution having a countably infinite number of arbitrarily small smooth pieces may occur.

Let $\{a,b\}$ and $\{c,d\}$ be rarefaction intervals of $f(u)$ as illustrated in the lower left hand side of Fig. 2.4. $f(u)$ has three inflection points in the interval $[d,b]$. Consider the corresponding generalized Riemann problem (2.8), (2.9). Using the construction methods of the proof of Theorem 2.3, it is seen that the generalized Oleĭnik solution initially has a single shock curve determined by $h_l(x)$ and $h_r(x)$ starting at x_0 with slope $dC(b,c;u)/du$. The rates at which the left and right states $u_l(t)$ and $u_r(t)$ change

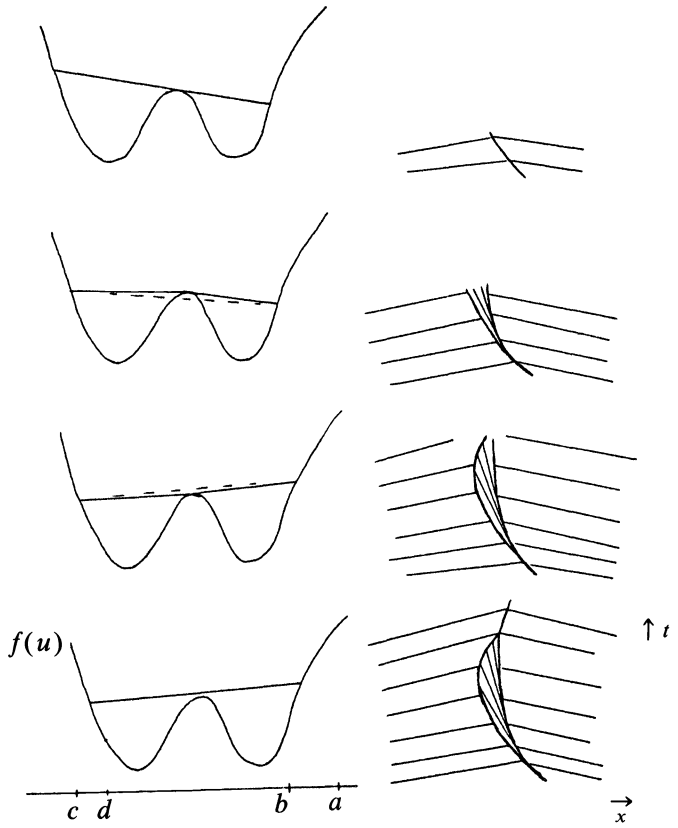


FIG. 2.4. *Formation of a single "bubble" in a solution. The dotted lines in the middle diagrams on the right emphasise the nonlinearity of the convex envelope.*

along the shock curve depend on the functions $h_l(x)$ and $h_r(x)$. Given $\{c,d\}$ (and hence $\{a,b\}$), $[x_0,x_r]$ and $h_r(x)$, it is possible to determine $h_l(x)$ and $[x_l,x_r]$ such that for all time $t>0$

$$E(u_l(t),u_r(t);u)=C(u_l(t),g(t);u)\cup C(g(t),u_r(t);u), \tag{2.24}$$

where

$$f'(g(t))=\frac{\partial C(u_l(t),g(t);u)}{\partial t}=\frac{\partial C(g(t),u_r(t);u)}{\partial t}. \tag{2.25}$$

The convex envelope (2.24), though consisting of two chords, defines a single shock line and represents a metastable balance between a single and a two shock configuration. This construction can be perturbed to create a single "bubble"; that is a time interval $t_1<t<t_2$ over which the shock separates into two shocks as is illustrated on the right in Fig. 2.4. This is accomplished by adjustment of the rate at which the rarefaction waves (i.e. adjustment of the respective h functions) enter the shock waves from the extreme left and right. The above construction can be iterated to produce a second bubble starting at t_3 by adjusting (slowing) the rate $h_l'(x)$ appropriately. It has not as yet been determined if, for a given set of intervals $\{a,b\}$, $\{c,d\}$, $[x_l,x_0]$, and $[x_0,x_r]$, it is possible to construct a sequence of generalized Riemann problems, where the ith problem is characterized by the choice of functions $h_l^i(x)$, $h_r^i(x)$, for which the solution to the ith problem has i such bubbles. If this was possible, the solution to the limiting problem obtained as $i\to\infty$ would not be piecewise smooth. We note it is certainly possible to create an infinite sequence of such problems if only $\{a,b\}$, $\{c,d\}$, and $[x_r,x_0]$ are held constant and the interval $[x_0,x_l]$ is allowed to vary with i. It remains to be proven that such a sequence of problems can be chosen such that $[x_0,x_l]$ also remains fixed.

3. The generalized Riemann problem for initial data with multiple points of discontinuity. Let $x_l<x_0<x_1<\cdots<x_n<x_r$. Consider (2.8) with initial data

$$u(0,x)=\begin{cases} u_l, & x<x_l,\\ h_l(x_l,x_0,u_l,u_0;x), & x_l<x<x_0,\\ h_j(x_j,x_{j+1},u_j,u_{j+1};x), & x_j<x<x_{j+1}, \quad j=0,1,\cdots,n-1,\\ h_r(x_n,x_r,u_n,u_r;x), & x_n<x<x_r,\\ u_r, & x_r\leqq x, \end{cases} \tag{3.1}$$

where $h_l(x)$, $h_j(x)$, $h_r(x)$ are as in (2.7), and $\{u_l,u_0\}$, $\{u_0,u_1\},\cdots,\{u_n,u_r\}$ are rarefaction intervals of f. The Cauchy problem (2.8), (3.1) is an extension of the definition of the generalized Riemann problem to include multiple initial discontinuities.

THEOREM 3.1. *The solution to* (2.8) *with initial data* (3.1) *for* $f\in C^2$: $\mathbf{R}\to\mathbf{R}$ *having at most one inflection point is piecewise smooth. There are no more than* $(n+1)!$ *shocks in the solution and no more than* $(n+1)!$ *irregular points.*

The restriction to either zero or one inflection point for $f(u)$ guarantees that when two shocks interact a single shock is produced. Thus given a finite number m of initial discontinuities, the number of shocks in the solution remains bounded by $m!$ and the solution will be piecewise smooth. For f having two or more inflection points there is no a priori bound on the growth of the number of shocks in time since an interaction of two (incoming) shocks can conceivably give rise to two or more (outgoing) shocks (or to

a single shock which later splits via interaction with rarefaction waves into two or more shocks) and the possibility of generating a solution that is richer in structure than piecewise smooth exists. It is only through the decrease in variation of the solution (ie. the decrease in shock strength with in time) that such a bound may be possible. Such a bound on the growth rate of shock number combined with a general statement for the >2 inflection point case of §2 is required for a more general statement on the piecewise smoothness of the solution to the generalized Riemann problem for initial data with multiple points of discontinuity.

Theorem 3.1 will be proven by finite induction on n in (3.1). The case $n=0$ is the generalized Riemann of §2.1. The case $n=1$ is given below. It suffices to illustrate the salient points of the general induction step.

Proof of Theorem 3.1 *for* $n=1$. Consider (2.8) under three possible cases of Cauchy data:

$$u(x)=\begin{cases} u_l, & x\leqq x_l,\\ h_l(x_l,x_0,u_l,u_0;x), & x_l<x<x_0,\\ h_0(x_0,x_1,u_0,u_1;x), & x_0<x<x_1,\\ u_1, & x_1\leqq x,\end{cases} \tag{3.2}$$

$$u(x)=\begin{cases} u_0, & x\leqq x_0,\\ h_0(x_0,x_1,u_0,u_1;x), & x_0<x<x_1,\\ h_r(x_1,x_r,u_1,u_r;x), & x_1<x<x_r,\\ u_r, & x_r\leqq x,\end{cases} \tag{3.3}$$

$$u(x)=\begin{cases} u_l, & x\leqq x_l,\\ h_l(x_l,x_0,u_l,u_0;x), & x_l<x<x_0,\\ h_0(x_0,x_1,u_0,u_1;x), & x_0<x<x_1,\\ h_r(x_1,x_r,u_1,u_r;x), & x_1<x<x_r,\\ u_r & x_r\leqq x.\end{cases} \tag{3.4}$$

From §2, construct the solutions to (3.2) and (3.3), denoting them $u_L(t,x)$ and $u_R(t,x)$ respectively. Denote the solution to (3.4) as $u_C(t,x)$. The following four (exhaustive but not mutually exclusive) possibilities exist for the solutions u_L and u_R:[4]

P1) There exists a rarefaction wave $(v,0,x_v)$ common to both solutions with $v\in\{u_0,u_1\}$, $x_v\in(x_0,x_1)$ which propagates unimpeded for all t.

P2) Both u_L and u_R have a shock curve and there exists a point $p_2=(t_2,x_2)$ common to the shock curve of each solution for some $t_2>0$.

P3) u_L has a shock curve, denoted $(u_l^1(t),u_r^1(t),x)$ and $u_r^1(t_3)=u_1$ for $t_3>0$.

P4) u_R has a shock curve, denoted $(u_l^2(t),u_r^2(t),x)$ and $u_r^2(t_4)=u_0$ for some $t_4>0$.

---

[4]As x_0 and x_1 are spatially separated, intuitively M_C is expected to be composed thusly:

$M_C=$ left piece of $M_L\cup$ piece due interaction of waves x_0 and $x_1\cup$ right piece of M_R.

Conditions P1 through P4 characterize the interactions between waves from x_0 and x_1.

If P1 holds (if neither solution has a shock wave, then condition P1 holds) then

$$u_C(x,t) \equiv \begin{cases} u_L(x,t) & \text{if } x \leqq x_v + f'(v)t, \\ u_R(x,t) & \text{if } x > x_v + f'(v)t, \end{cases} \tag{3.5}$$

is the unique solution to (2.8) (3.1) obeying (1.4) and (1.5).

If P1 does not hold let $t_{\min}$ denote the minimum of the existing values t_2, t_3, t_4.

If $t_{\min} = t_2$ then let $\Sigma_L^0 \equiv (u_l^0(t), u_r^0(t), 0, x_0)$ and $\Sigma_R^1 \equiv (u_l^1(t), u_r^1(t), 0, x_1)$ denote the shock waves propagating respectively from x_0 in u_L and x_1 in u_R. Then for $t \leqq t_2$

$$u_C(x,t) \equiv \begin{cases} u_L(x,t) & \text{if } x \leqq h_0^{-1}(v_2) + f'(v_2)t, \\ u_R(x,t) & \text{if } x > h_0^{-1}(v_2) + f'(v_2)t, \end{cases} \tag{3.6}$$

where $v_2 = u_r^0(t_2) = u_l^1(t_2)$ is in the interval $\{u_0, u_1\}$. For $t > t_2$, $u_C(x,t)$ is the generalized Oleĭnik solution to (2.8) with initial data (at t_2)

$$u(t_2,x) = \begin{cases} u_l, & x \leqq x_l^{t_2}, \\ h_l^{t_2}(x_l^{t_2}, x_2, u_l, u_l^1(t_2); x), & x_l^{t_2} < x < x_2, \\ h_r^{t_2}(x_2, x_r^{t_2}, u_r^2(t_2), u_r; x), & x_2 < x < x_r^{t_2}, \\ u_r, & x_r^{t_2} \leqq x, \end{cases}$$

where

$$x_l^{t_2} = x_l + f'(u_l)t_2, \qquad x_r^{t_2} = x_r + f'(u_r)t_2.$$

$h_l^{t_2}(x)$ is defined by the rarefaction interval $\{u_l, u_l^1(t_2)\}$ where $u_l^1(t_2)$ is the left limit of u at (t_2, x_2) for the shock curve Σ_0 in u_C propagating from x_0 in (3.6). (Σ_0 is identical with Σ_L^0 for $0 \leqq t \leqq t_2$.) Similarly $h_r^{t_2}(x)$ is defined by the rarefaction interval $\{u_r, u_r^2(t_2)\}$ where $u_r^2(t_2)$ is the right limit of u at (t_2, x_2) for the shock curve Σ_1 propagating from x_1. (Σ_1 is identical with Σ_L^1 for $0 \leqq t \leqq t_2$.)

If $t_{\min} = t_3$, let $x_3 = x_1 + f'(u_1)t$. There exist two possibilities for the solution u_C if u_R has a shock wave. (If u_R has no shock the solution for u_C is deduced easily from the discussion below.) These two possibilities are either that u_R has a leftmost rarefaction fan

a) centered at the point x_1 (Fig. 3.1a),

b) determined by the shock wave Σ_1 originating at $(0, x_1)$ (Fig. 3.1b).

We note that in either case this leftmost rarefaction fan in u_R is a smooth continuation of the rightmost rarefaction fan in the solution u_L for $t \leqq t_3$. For $t \leqq t_3$, let

$$u_C(t,x) = \begin{cases} u_L(t,x), & x \leqq x_1 + f'(u_1)t, \\ u_R(t,x), & x \geqq x_1 + f'(u_1)t. \end{cases} \tag{3.7}$$

At (t_3, x_3) the shock Σ_0 originating from $(0, x_0)$ encounters the first rarefaction wave from the solution for $x > x_3$. For case a) above, the shock Σ_0 is determined for $t > t_3$ by the rarefaction fan determined by $h_l(x_l, x_0, u_l, u_0; x)$ and the rarefaction fan centered on $(0, x_1)$. Denote the curve followed by the shock Σ_0 for $t > t_3$ as $(t, x_{\Sigma_0}(t))$. Then, for case a),

$$u_C(t,x) = \begin{cases} u_L, & x < x_{\Sigma_0}(t), \\ u_R, & x > x_{\Sigma_0}(t), \end{cases} \qquad t > t_3. \tag{3.8}$$

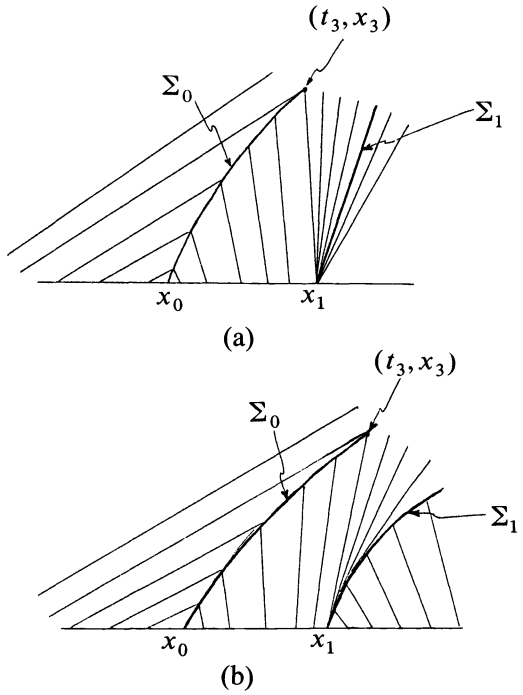


FIG. 3.1. *Two possibilities encountered under case* P3 *described in the text. Dark lines are shock waves, light lines are rarefaction waves.*

In case b) the shock Σ_0 is determined for $t > t_3$ by the rarefaction fan determined by $h_l(x_l, x_0, u_l, u_0; x)$ and the rarefaction fan determined by the shock line Σ_1. In either case a) or b), the resulting shock, Σ_0 hits the shock Σ_1 at some time $t_5 > t_3$ or it does not. If it does not, the solution for all $t > t_3$ is given by (3.8). If it does, the solution for $t_3 \leqq t \leqq t_5$ is given by (3.8) and for $t \geqq t_5$ by the solution to a generalized Riemann problem of the type discussed in §2.1. It is to be noted that the solution is C^1 continuous at t_3 (ie. all waves propagate continuously from $t_3 - \varepsilon$ to $t_3 + \varepsilon$ and all waves except the interacting shock lines are C^1 continuous at t_5.

If $t_{\min} = t_4$, an analogous construction to that for case P3 holds.

In all cases the constructed solution u_C is thus piecewise smooth. The general induction step for Theorem 3.1 follows the above argument with a more complicated notation. □

4. The two-dimensional Riemann problem. For convenience we shall use the notation $f_1 \equiv f$, $f_2 \equiv g$, $x_1 \equiv x$, $x_2 \equiv y$ for the discussion in two spatial dimensions.

DEFINITION 4.1. The Cauchy problem (1.1) with initial data piecewise constant on a finite number of wedges focused on a single point in the x, y plane shall be referred to as the *two-dimensional Riemann problem*. Without loss of generality, this point can be taken to be the point $x = 0$, $y = 0$.

For the case $g \equiv f$, under the 45° rotation $2\xi = x + y$, $2\eta = y - x$, (1.1) becomes

$$\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial \xi} = 0. \tag{4.1}$$

From (4.1) we see that the solution can be obtained along each $\eta = constant$ plane independent of other η. In particular, this implies the solutions obtained in the $\eta < 0$ half space can be obtained independently of the solutions in the $\eta > 0$ half; by the symmetry of the problem, no new solutions will be found in the $\eta < 0$ half space that are not found in the upper. We therefore restrict our discussion to the half space $\eta > 0$ ($y > x$). The solution in the plane $\eta = 0$ is discussed in Lemma 4.4.

In the $\eta > 0$ half space, the solution in each $\eta = constant$ plane is just that for a generalized Riemann problem in one dimension.

LEMMA 4.2. *Let $u(t,\xi,\eta_0)$ denote the generalized Riemann problem solution in the $\eta = \eta_0 > 0$ plane. Then the solution $u(t,\xi,\eta_1)$ to the generalized Riemann problem in the η_1 plane ($\eta_1 > 0$) is just a similarity solution of $u(t,\xi,\eta_0)$ given by*

$$u(ct, c\xi + b(c), \eta_1) = u(t,\xi,\eta_0), \tag{4.2}$$

where $c = \eta_1/\eta_0$.

Proof. If $v_0(\xi)$ is the initial data for the generalized Riemann problem on η_0, then $w_0(\zeta \equiv c\xi + b(c)) = v_0(\xi)$ is the initial data for the generalized Riemann problem on η_1 where $c = \eta_1/\eta_0$ and $b(c)$ is given by the wedge pattern of the two-dimensional Riemann problem. Further, the function $u(\tau \equiv ct, \zeta \equiv c\xi + b, \eta_1 \equiv c\eta_0) \equiv u(t,\xi,\eta_0)$ is a weak solution of the generalized Riemann problem on η_1 having initial data w_0. □

Thus the solution to the two-dimensional Riemann problem (4.1) in the $\eta > 0$ half space is found from a mapping M: $\mathbf{R}^1 \to \mathbf{R}^2$ which is similarity transformation in t,ξ,η and therefore continuous onto the domain $t > 0$, ξ, $\eta > 0$.

THEOREM 4.3. *The unique (in the sense of Kružkov) solution in the plane $\eta > 0$ ($y > x$), to (1.1) with initial data that is piecewise constant on a finite number of wedges focused on a single point in the plane, with $f_1 \equiv f_2 \equiv f$, $f \in C^2$: $\mathbf{R} \to \mathbf{R}$, f having at most one inflection point, is*

a) *piecewise smooth,*

b) *composed of nonlinear waves and constant states which are the images under a similarity map M defined by Lemma 4.2 of the rarefaction and shock waves and constant states of a generalized Riemann problem in one dimension,*

c) *composed of curves of irregular points corresponding to images under the map M of the irregular points in the one-dimensional Riemann problem.*

Proof. Follows immediately from Lemma 4.2 and Theorem 3.1. □

LEMMA 4.4. a) *If the half line $\eta = 0$, $\xi < 0$ is not a line of discontinuity of the initial data (i.e., is not a wedge line) the waves incident upon the corresponding half plane $t > 0$, $\xi < 0$, $\eta = 0$ are continuous across the half plane.*

b) *If the half line $\eta = 0$, $\xi < 0$ is a line of discontinuity (constant jump) of the initial data, it remains a half plane $t > 0$, $\xi < 0$, $\eta = 0$ of (in general variable) jump discontinuity in the solution.*

Proof. a) Follows from taking the limit $\eta \to 0$ of the solution in either half plane.

b) Follows from taking the limit of solutions for which the wedge line in the initial data is displaced slightly from the $\eta = 0$ axis. □

Similar statements hold for the half line $\eta = 0$, $\xi > 0$.

Although the map M provides a means of constructing the solution to the two-dimensional Riemann problem, it is not a convenient method of doing so. In particular it is a method that will not generalize to cases $f \neq g$ where no appeal to a one-dimensional analysis can be made. We propose that the correct method of dealing with the general solution to the two-dimensional Riemann problem is to identify the general

two-dimensional nonlinear waves. In [4] we identify these waves for the case $f \equiv g$ and use the construction method initiated by Guckenheimer [1] and Wagner [6] to piece these waves together into entropy obeying solutions.

5. Concluding remarks. The condition $f_1 \equiv f_2 \equiv f$, $f \in C^2$: $\mathbf{R} \to \mathbf{R}$, f having at most one inflection point has been shown to be sufficient to guarantee that the solution to the two-dimensional Riemann problem is piecewise smooth. A construction has been presented detailing a presumed mechanism for loss of piecewise smoothness if f has three or more inflection points.

The Riemann problem in one dimension has been generalized to include a finite number of discontinuities and smooth "rarefaction" variation in the initial data. Sufficiency conditions based upon the number of inflection points in the convection function f have been obtained which guarantee piecewise smoothness of the solution to the generalized problem.

Acknowledgments. The author wishes to thank Professor James Glimm for suggesting this problem and for his guidance on several issues of this paper. Thanks also to Jonathan Goodman for several helpful discussions.

REFERENCES

[1] J. GUCKENHEIMER, *Shocks and rarefactions in two space dimensions*, Arch. Rational Mech. Anal., 59 (1975), pp. 281–291.

[2] S. N. KRUŽKOV, *First order quasilinear equations in several independent variables*, Mat. USSR-Sb, 10 (1970), pp. 217–243.

[3] P. D. LAX, *Hyperbolic Systems of Conservation Laws and the Mathematical Theory of Shock Waves*, CBMS Regional Conference Series in Applied Mathematics 11, Society for Industrial and Applied Mathematics, Philadelphia, 1973.

[4] W. B. LINDQUIST, *Construction of solutions for two-dimensional Riemann problems*, DOE Research and Development Report DOE/ER/03077-228, Sept. 1984.

[5] O. A. OLEĬNIK, *Uniqueness and stability of the generalized solution of the Cauchy problem for a quasi-linear equation*, Uspekhi Mat. Nuak, 14 (1959), pp. 165–170, English trans., Amer. Math. Soc. Transl. Ser. 2, 33 (1963), pp. 285–290.

[6] D. WAGNER, *The Riemann problem in two space dimensions for a single conservation law*, this Journal, 14 (1983), pp. 534–559.

SIAM J. MATH. ANAL.
Vol. 17, No. 5, September 1986

010

ON THE IDENTIFICATION OF PARAMETERS IN GENERAL VARIATIONAL INEQUALITIES BY ASYMPTOTIC REGULARIZATION*

K.-H. HOFFMANN† AND J. SPREKELS†

Abstract. A method for the identification of parameters in a general class of variational inequalities from the knowledge of the solutions is proposed. The method consists of embedding the original problem into a sequence of regularizing equations. An a priori estimate is fundamental to prove that limit points of the regularizing sequence are solutions of the original problem. The method is applied to various applications, in particular, to the dam problem and to linear elasticity with friction.

Key words. distributed parameter identification, variational inequalities, regularization, inverse problems

AMS(MOS) subject classifications. Primary 35R30, 65M30

1. Introduction. Starting point for the present paper was the following identification problem which is considered in [3]:

(1.1) Given $u^* \in \mathring{H}^1(\Omega) \cap H^{1,\infty}(\Omega)$ and $f^* \in H^{-1}(\Omega)$, find a matrix $a^* = (a_{ij}^*)$ with entries $a_{ij}^* \in L^\infty(\Omega)$, such that $-\nabla \cdot (a^* \nabla u^*) = f^*$.

This problem is closely related to a problem which occurs in oil field exploration (see [4]). In [3] a method for the solution of (1.1) was proposed which consists in trying to obtain (1.1) as asymptotic steady-state of a regularizing system of approximating parameter-dependent problems. The basic idea was to construct the regularized equations in such a way that a Lyapunov-type a priori estimate holds from which convergence results can be derived.

In [3] this technique was specified and applied numerically. A slightly different regularization was used in [6]. In [7] the above idea was applied to compartment models for ordinary differential equations.

In this paper we give a general functional analytic framework for the technique under which the results of [3], [6], [7] can be subsumed. Moreover, we show that the method may be used for identifying the system coefficients of various other problems involving partial differential equations, among those the problem of identifying the permeability matrix in the dam problem from measurements of the pressure and the problem of the coefficients of elasticity in linear elasticity from measurements of the displacements.

The paper is organized as follows. In §2 our method is formulated in functional analytic framework and is applied to general elliptic variational inequalities. Convergence and stability results are proved. The method used here differs slightly from that used in the previous papers [3], [6], [7]. In §3 we extend some of the results of §2 to the problem of identifying parameters in evolution equations. The concluding §4 brings various applications.

2. Stationary variational inequalities. We now define our general functional analytic setting. Let the separable and reflexive Banach space V be dense with continuous

*Received by the editors February 16, 1984, and in revised form February 14, 1985.
† Universität Augsburg, Memminger Strasse 6, 8900 Augsburg, West Germany.

embedding in the Hilbert space H. Then $V \subset H \subset V^*$ with dense and continuous embedding where V^* is the dual of V. The dual pairing between elements of V^* and V is denoted by $\langle \cdot, \cdot \rangle$. By C we denote some nonempty, closed and convex subset of V. Moreover, let X denote another Hilbert space with the inner product $[\cdot, \cdot]$. The norms of the above spaces are denoted by $\|\cdot\|_V$, $\|\cdot\|_H$, $\|\cdot\|_{V^*}$, $\|\cdot\|_X$. Finally, we assume that $(X_0, \|\cdot\|_{X_0})$ is another Banach space such that X_0 is a dense subspace of X.

We then consider the parameter identification problem:

(P) Given $u^* \in D(S) \cap C$ and $f^* \in V^*$, find $a^* \in X$ with $(a^*, u^*) \in D(A_2)$ such that there is a $w^* \in S(u^*)$ which satisfies the variational inequality

$$\langle w^* + A_1(a^*) + A_2(a^*, u^*) - f^*, v - u^* \rangle + \Psi(v) - \Psi(u^*) \geqq 0 \quad \forall v \in C. \tag{2.1}$$

Here and throughout we assume:

(A1) $S : D(S) \subset V \to 2^{V^*}$ is a maximal monotone graph with $u^* \in \operatorname{int} D(S)$.

(A2) $A_1 : X \to V^*$ is linear and bounded.

(A3) $A_2 : D(A_2) \subset X \times V \to V^*$ is a bilinear map such that

(i) $(X_0 \times V) \subset D(A_2)$,

(ii) $|\langle A_2(a,u), v \rangle| \leqq \gamma_1 \|a\|_{X_0} \|u\|_V \|v\|_V$, $\forall a \in X_0$, $u, v \in V$, for some $\gamma_1 > 0$.

(A4) $\Psi : V \to \mathbb{R}$ is continuous and convex.

Remark 1. $u^* \in \operatorname{int} D(S)$ implies that S is locally bounded at u^* (see [8, p. 31, Lemma 2.3]).

Remark 2. From (A3ii) follows that $A_2(a_n, u_n) \to A_2(a, u)$ weakly in V^* if $(a_n, u_n) \in X_0 \times V$ and $\|a_n - a\|_{X_0} \to 0$, $\|u_n - u\|_V \to 0$.

In §4 the above assumptions are verified for various applications. We assume:

(A5) (P) has at least one solution $a^* \in X_0$ such that $\langle A_2(a^*, u - v), u - v \rangle \geqq \delta \|u - v\|_V^2$, $\forall u, v \in C$, for some $\delta > 0$ (uniform monotonicity on C).

We now introduce a sequence of systems of variational inequalities the solutions of which shall for $n \to \infty$ converge to a solution of a finite-dimensional analogue of (P).

To this end, let V_N and W_M be finite-dimensional subspaces of V and X_0 such that:

(A6) $u^* \in V_N$.

(A7) $\langle A_1(a) + A_2(a, u), v \rangle = 0$, whenever $a \in W_M^{\perp}$, $u \in C$, and $v \in C - C$. Here $W_M^{\perp} := \{ z \in X_0 : [z, \eta] = 0 \ \forall \eta \in W_M \}$.

Whereas (A6) is easily verified, the compatibility condition (A7) is not obvious. We refer to §4.

Now let $\varepsilon > 0$ be arbitrary. For fixed $h > 0$ we consider the problem:

(P_h) Given $(a_0, u_0) \in W_M \times (D(S) \cap C \cap V_N)$, find $(a_n, u_n) \in W_M \times (D(S) \cap C \cap V_N)$ such that there exists $w_n \in S(u_n)$ $(n \in \mathbb{N})$ with

$$\left\langle \varepsilon \frac{u_{n+1} - u_n}{h} + w_{n+1} + A_1(a_{n+1}) + A_2(a_{n+1}, u_{n+1}) - f^*, v - u_{n+1} \right\rangle + \Psi(v) - \Psi(u_{n+1}) \geqq 0 \quad \forall v \in C \cap V_N, \tag{2.2}$$

$$\left[\frac{a_{n+1} - a_n}{h}, \eta \right] = \langle A_1(\eta) + A_2(\eta, u_{n+1}), u_{n+1} - u^* \rangle \quad \forall \eta \in W_M. \tag{2.3}$$

Let us postpone the question of solvability and assume that $\{(a_n, u_n)\}_{n \geqq 0}$ solves (P_h). We derive an a priori estimate which is fundamental for all subsequent considerations.

To this end, put $q_n := u_n - u^*$, $r_n := a_n - a^*$.

Substitution of $v := u^*$ into (2.2) and $v := u_{n+1}$ into (2.1) gives:

$$\left\langle \varepsilon \frac{q_{n+1}-q_n}{h} + w_{n+1} + A_1(a_{n+1}) + A_2(a_{n+1}, u_{n+1}) - f^*,\ q_{n+1} \right\rangle + \Psi(u_{n+1}) - \Psi(u^*)$$

$$\leqq 0 \leqq \Psi(u_{n+1}) - \Psi(u^*) + \left\langle w^* + A_1(a^*) + A_2(a^*, u^*) - f^*, q_{n+1} \right\rangle,$$

whence

$$(2.4) \quad 0 \geqq \left\langle \varepsilon \frac{q_{n+1}-q_n}{h},\ q_{n+1} \right\rangle + \delta \|q_{n+1}\|_V^2 + \left\langle A_1(r_{n+1}) + A_2(r_{n+1}, u_{n+1}),\ q_{n+1} \right\rangle,$$

where the monotonicity of S and (A5) were used. Let P denote the $[\cdot,\cdot]$-orthogonal projection onto the closed subspace W_M of X. Substitution of $\eta := a_{n+1} - Pa^*$ into (2.3) yields by (A7) and $a^* - Pa^* \in W_M^{\perp}$:

$$\begin{aligned} 0 &= \left[\frac{a_{n+1}-a_n}{h},\ r_{n+1} \right] + \left[\frac{a_{n+1}-a_n}{h},\ a^* - Pa^* \right] \\ &\quad - \left\langle A_1(r_{n+1}) + A_2(r_{n+1}, u_{n+1}),\ q_{n+1} \right\rangle \\ &\quad + \left\langle A_1(Pa^* - a^*) + A_2(Pa^* - a^*, u_{n+1}),\ q_{n+1} \right\rangle \\ &= \left[\frac{r_{n+1}-r_n}{h},\ r_{n+1} \right] - \left\langle A_1(r_{n+1}) + A_2(r_{n+1}, u_{n+1}) q_{n+1} \right\rangle. \end{aligned}$$

Addition to (2.4) gives:

$$0 \geqq \left\langle \varepsilon \frac{q_{n+1}-q_n}{h},\ q_{n+1} \right\rangle + \left[\frac{r_{n+1}-r_n}{h},\ r_{n+1} \right] + \delta \|q_{n+1}\|_V^2.$$

Hence, from standard arguments, for all $n \geqq 0$,

$$(2.5) \qquad \varepsilon \|q_{n+1}\|_H^2 + \|r_{n+1}\|_X^2 + 2\delta h \|q_{n+1}\|_V^2 \leqq \varepsilon \|q_n\|_H^2 + \|r_n\|_X^2,$$

and upon summation,

$$(2.6) \qquad \frac{\varepsilon}{2h} \|q_{n+1}\|_H^2 + \frac{1}{2h} \|r_{n+1}\|_X^2 + \delta \sum_{k=0}^{n} \|q_{k+1}\|_V^2 \leqq \frac{\varepsilon}{2h} \|q_0\|_H^2 + \frac{1}{2h} \|r_0\|_X^2.$$

Hence we have proved the a priori estimate

$$(2.7) \qquad \sup_{n \in \mathbb{N}} \left[\varepsilon \|q_n\|_H^2 + \|r_n\|_X^2 \right] + 2\delta h \sum_{k=1}^{\infty} \|q_k\|_V^2 \leqq C < \infty,$$

where C depends on ε, a_0, u_0, u^*, a^*, but not on h, V_N, W_M.

Remark 3. Note that (2.7) holds for any solution $a^* \in X_0$ such that $A_2(a^*, \cdot)$ is uniformly monotone on $C \cap V_N$.

THEOREM 2.1. *Let the assumptions* (A1)–(A7) *hold. Then there is an* $h_0 > 0$ *such that* (P_h) *has a solution* $\{(a_n, u_n)\}_{n \geqq 0}$ *whenever* $0 < h \leqq h_0$.

The proof uses the following general result on variational inequalities (see [8, p. 197, Satz 2.7]).

THEOREM 2.2. *Let* B *be a reflexive Banach space and* $K \subset B$, $K \neq \emptyset$, *be closed and convex. Let* $g : K \to \mathbb{R}$ *be convex and lower semicontinuous. Moreover, let* $P_1 : D(P_1) \subset B \to 2^{B^*}$ *be maximal monotone, and let* $P_2 : K \to B^*$ *be continuous, bounded, pseudomonotone and coercive with respect to* $\Theta \in B^*$. *Moreover, let the following conditions hold*:

$$(2.8) \qquad \operatorname{int} D(P_1) \cap D(\partial g) \neq \emptyset,$$

$$\partial(g+\chi_K)=\partial g+\partial\chi_K, \tag{2.9}$$

$$K\cap \operatorname{int} D(P_1+\partial g)\neq\varnothing. \tag{2.10}$$

Then there is a $u\in K$ *such that for some* $w\in P_1(u)$:

$$\langle w+P_2(u),\, v-u\rangle+g(v)-g(u)\geqq 0 \quad \forall v\in K. \tag{2.11}$$

(*The terms bounded, pseudomonotone, coercive have the usual meaning*; *cf.* [8]; χ_K *is the characteristic function of* K *and* ∂ *the subgradient*).

Proof of Theorem 2.1. Let $W_M := \operatorname{span}\{\Psi_1,\cdots,\Psi_M\}$ with $[\Psi_i,\Psi_j]=\delta_{ij}$, $\forall i,\ j$. We conclude inductively. So let (a_k,u_k), $0\leqq k\leqq n$, already be constructed. We consider:

$$\left[\frac{a-a_n}{h},\,\eta\right]-\langle A_1(\eta)+A_2(\eta,u),\, u-u^*\rangle=0 \quad \forall\eta\in W_M. \tag{2.12}$$

Given $u\in C$, (2.12) has the unique solution

$$a=a_n+h\sum_{k=1}^{M}\langle A_1(\Psi_k)+A_2(\Psi_k,u),\, u-u^*\rangle\Psi_k. \tag{2.13}$$

Hence the solution of (2.2), (2.3) is equivalent to finding $u\in D(S)\cap C\cap V_N$ such that with some $w\in S(u)$:

(2.14)

$$\left\langle \varepsilon\frac{u-u_n}{h}+w+A_1(a)+A_2(a,u)-f^*,\, v-u\right\rangle+\Psi(v)-\Psi(u)\geqq 0 \quad \forall v\in C\cap V_N,$$

with a given by (2.13).

We apply Theorem 2.2 with the following specifications:

$$B=V,\quad K=C\cap V_N,\quad P_1=S,\quad g(v)=\Psi(v)-\left\langle\frac{\varepsilon}{h}u_n+f^*,\, v\right\rangle,$$

$$P_2(u):=\frac{\varepsilon}{h}u+A_1(a_n)+h\sum_{k=1}^{M}\langle A_1(\Psi_k)+A_2(\Psi_k,u),\, u-u^*\rangle A_1(\Psi_k)$$

$$+A_2(a_n,u)+h\sum_{k=1}^{M}\langle A_1(\Psi_k)+A_2(\Psi_k,u),\, u-u^*\rangle A_2(\Psi_k,u).$$

K is closed and convex, and $u^*\in K$. P_1 is maximal monotone. By (A1), $u^*\in\operatorname{int} D(P_1)$ and $u^*\in D(\partial g)$ from the continuity of Ψ which yield (2.8). Moreover g and χ_K are proper convex functionals, and g is continuous at u^* which lies in the effective domains of g and χ_K. Hence (2.9) holds (see [8, p. 125, Satz 5.16]). Finally, $V=D(\partial g)$ and thus $D(P_1+\partial g)=D(P_1)$. Hence (2.10) follows from $u^*\in K\cap\operatorname{int} D(P_1)$.

It remains to show that P_2 has the required properties.

α) $P_2: K\to V^*$ is bounded: This follows from (A3ii) and:

$$\|P_2(u)\|_{V^*}\leqq\frac{\varepsilon}{h}\|u\|_{V^*}+\|A_1(a_n)\|_V+\gamma_1\|a_n\|_{X_0}\|u\|_V$$

$$+h\sum_{k=1}^{M}\left\{\|A_1(\Psi_k)\|_{V^*}+\gamma_1\|\Psi_k\|_{X_0}\|u\|_V\right\}$$

$$\cdot\|u-u^*\|_V\left\{\gamma_1\|\Psi_k\|_{X_0}\|u\|_V+\|A_1(\Psi_k)\|_{V^*}\right\}.$$

β) $P_2: K \to V^*$ is continuous: This follows from a lengthy but straightforward computation which uses the bilinearity of A_2 and (A3ii) as essential tools.

γ) $P_2: K \to V^*$ is pseudomonotone: Let $\{u_m\} \subset K$ be given with $u_m \to u$ weakly in V and $\limsup \langle P_2(u_m), u_m - u \rangle \leqq 0$. From $\dim W_M < \infty$ it follows that $u_m \to u$ strongly in V and hence by β): $P_2(u_m) \to P_2(u)$ strongly in V^*. Hence for all $w \in V$:

$$\langle P_2(u), u - w \rangle \leqq \liminf \langle P_2(u_m), u_m - w \rangle .$$

δ) $P_2: K \to V^*$ is coercive with respect to $\Theta \in B^*$ whenever $0 < h \leqq h_0$, where $h_0 > 0$ does not depend on n: Let $u \in K$. We then have by (A3ii):

$$\begin{aligned}
\langle P_2(u), u \rangle &= \frac{\varepsilon}{h} \|u\|_H^2 + \langle A_1(a_n), u \rangle + \langle A_2(a_n, u), u \rangle \\
&\quad + h \sum_{k=1}^{M} \langle A_1(\Psi_k) + A_2(\Psi_k, u), u - u^* \rangle \langle A_1(\Psi_k) + A_2(\Psi_k, u), u \rangle \\
&\geqq \frac{\varepsilon}{h} \|u\|_H^2 - \|A_1(a_n)\|_{V^*} \|u\|_V - \gamma_1 \|a_n\|_{X_0} \|u\|_V^2 \\
&\quad + h \sum_{k=1}^{M} \left[\langle A_1(\Psi_k) + A_2(\Psi_k, u), u \rangle - \frac{1}{2} \langle A_1(\Psi_k) + A_2(\Psi_k, u), u^* \rangle \right]^2 \\
&\quad - \frac{h}{4} \sum_{k=1}^{M} \langle A_1(\Psi_k) + A_2(\Psi_k, u), u^* \rangle^2 .
\end{aligned}$$

Since (a_n, u_n) solves (2.2), (2.3), the a priori estimate (2.7) yields that $\{a_n\}$ is bounded in X by a constant which does not depend on n. This together with the finite dimension of W_M, V_N, the boundedness of A_1 and (A3ii) yields:

$$\begin{aligned}
\langle P_2(u), u \rangle &\geqq \left(C_1 \frac{\varepsilon}{h} - \gamma_1 C_2 - \frac{h}{4} \gamma_1^2 \sum_{k=1}^{M} \|\Psi_k\|_{X_0}^2 \|u^*\|_V^2 \right) \|u\|_V^2 \\
&\quad - \left(C_3 + \frac{h}{2} C_4 \sum_{k=1}^{M} \gamma_1 \|\Psi_k\|_{X_0}^2 \|u^*\|_V^2 \right) \|u\|_V - \frac{h}{4} \sum_{k=1}^{M} \langle A_1(\Psi_k), u^* \rangle^2 .
\end{aligned}$$

Now choose $h_0 > 0$ so small that the expression in the first bracket becomes positive, which is possible independently of n. Then it follows that $\langle P_2(u), u \rangle > 0$ for $\|u\|_V \geqq R$, R sufficiently large, which proves the coercivity. This concludes the proof. □

THEOREM 2.3. *Let* (A1)–(A7) *hold, and let* $0 < h \leqq h_0$ ($h_0 > 0$ *as in Theorem* 2.1). *Let* $\{(a_n, u_n)\}_{n \geqq 0}$ *solve* (P_h). *Then it follows that*:

(i) $u_n \to u^*$, *strongly in* V.

(ii) *Every subsequence of* $\{a_n\}$ *has a cluster point* $a_\infty \in W_M$ *such that with some* $w_\infty \in S(u^*)$

(2.15)

$$\langle w_\infty + A_1(a_\infty) + A_2(a_\infty, u^*) - f^*, v - u^* \rangle + \Psi(v) - \Psi(u^*) \geqq 0 \quad \forall v \in C \cap V_N$$

(*i.e.*, a_∞ *solves the finite dimensional analogue of* (P)).

(iii) *If* $\Psi \equiv 0$, $C = V$, $S(u^*) = \{w_\infty\}$, *and* $\|u - v\|_H \leq \text{const} \cdot \|u - v\|_V \ \forall u, v \in V$ *then the whole sequence* $\{a_n\}$ *converges to some* $a_\infty \in W_M$ *such that* (2.15) *holds.*

Proof. Equation (2.7) implies (i). Moreover $\{a_n\}$, and hence any subsequence $\{a_m\}$, is bounded in X and has a weak cluster point $a_\infty \in X$. From $\dim W_M < \infty$ it follows without loss of generality that $\|a_m - a_\infty\|_{X_0} \to 0$ and thus $a_\infty \in W_M$.

By (A1) S is locally bounded at u^* (compare Remark 1). Since $\|u_m - u^*\|_V \to 0$, we may assume that $w_m \to w_\infty \in V^*$ weakly-star, and by reflexivity of V, weakly. The maximal monotonicity of S implies $w_\infty \in S(u^*)$ (compare [8, p. 33, Lemma 2.12]).

We show the validity of (2.15). To this end, consider (2.2). From $\|u_m - u^*\|_V \to 0$ and $\|a_m - a_\infty\|_{X_0} \to 0$ it follows by Remark 2 that $A_2(a_m, u_m) \to A_2(a, u)$ weakly in V^*. By (A2), $A_1 : X \to V^*$ is continuous, and thus $A_1(a_m) \to A_1(a_\infty)$ strongly in V^*. Moreover, $\Psi(u_{m+1}) \to \Psi(u^*)$, by (A4). Passing to the limit as $m \to \infty$ in (2.2) gives (2.15).

It remains to show (iii). Let $\Psi \equiv 0$, $C = V$ and $S(u^*) = \{w_\infty\}$. Then (2.1) is equivalent to the equation

$$w_\infty + A_1(a^*) + A_2(a^*, u^*) - f^* = \Theta, \tag{2.16}$$

and (2.15) reads

$$\left\langle w_\infty + A_1(a_\infty) + A_2(a_\infty, u^*) - f^*,\ v \right\rangle = 0 \quad \forall v \in V_N. \tag{2.17}$$

The solution set $L(V_N) := \{a_\infty \in X_0$: (2.17) holds$\}$ is an affine subspace of X_0, and $a^* \in L(V_N)$. Moreover, for every $\hat{a} \in L(V_N)$ such that $A_2(\hat{a}, \cdot)$ is uniformly monotone on $C \cap V_N$, the a priori estimate (2.7) is valid (compare Remark 3).

Now let $\hat{a} \in L(V_N)$ be arbitrary where $A_2(\hat{a}, \cdot)$ is uniformly monotone on $C \cap V_N$. Then by (2.5) there exists

$$l := \lim_{n \to \infty} \left(\varepsilon \|u_n - u^*\|_H^2 + \|a_n - \hat{a}\|_X^2 \right). \tag{2.18}$$

But $\|u_n - u^*\|_V^2 \to 0$, and thus $\|u_n - u^*\|_H^2 \to 0$, i.e.,

$$l = \lim_{n \to \infty} \|a_n - \hat{a}\|_X^2. \tag{2.19}$$

Now let $a_\infty^1, a_\infty^2 \in W_M$ be any two limit points of $\{a_n\}$. Then

$$\|a_\infty^1 - \hat{a}\|_X^2 = \|a_\infty^2 - \hat{a}\|_X^2. \tag{2.20}$$

For $\hat{a} := a^* + \lambda(a_\infty^1 - a_\infty^2) \in L(V_N)$ and sufficiently small $|\lambda| > 0$ the operator $A_2(\hat{a}, \cdot)$ is uniformly monotone on $C \cap V_N$. Hence

$$\|a_\infty^1 - a^* - \lambda(a_\infty^1 - a_\infty^2)\|_X^2 = \|a_\infty^2 - a^* - \lambda(a_\infty^1 - a_\infty^2)\|_X^2 \quad \text{or} \quad \|a_\infty^1 - a_\infty^2\|_X^2 = 0.$$

Thus $\{a_n\}$ has a unique limit point for $n \to \infty$. □

Remark 4. For numerical purposes it is noteworthy that the convergence of $\|a_n - a_\infty\|_{X_0}$ can be accelerated by a proper choice of the scaling factor ε.

Remark 5. Under the conditions of (iii) we have form $u^* \in V_N$:

$$\begin{aligned} \left\langle A_1(a_\infty) + A_2(a_\infty, u^*),\ u^* \right\rangle &= \left\langle f^* - w_\infty,\ u^* \right\rangle \\ &= \left\langle A_1(a^*) + A_2(a^*, u^*),\ u^* \right\rangle. \end{aligned}$$

Hence if $A_1 = \Theta$ and $u^* \neq \Theta$, then $\left\langle A_2(a_\infty, u^*),\ u^* \right\rangle > 0$ which means that $A_2(a_\infty, \cdot)$ is positive definite at u^*.

Next we show that limit points a_∞ of (P_h) are stable against perturbations in the initial data if $A_2(a_\infty, \cdot)$ is uniformly monotone on C and a_∞ is unique. However, for technical reasons we have to impose the assumption $A_1 = \Theta$. Also, the admissible set of parameters h has to be restricted.

THEOREM 2.4. *Let* (A1)–(A7) *hold with* $A_1 = \Theta$. *Assume initial data* $(a_0^\lambda, u_0^\lambda)$, $|\lambda| \leqq \lambda_0$, *are given such that*

$$a_0^\lambda \to a_0^0, \qquad u_0^\lambda \to u_0^0 \quad \text{as } \lambda \to 0. \tag{2.21}$$

Then there is an $\hat{h} \in (0, h_0]$ *such that for* $0 < h \leqq \hat{h}$ *the following assertion holds: If* $\{a_n^\lambda, u_n^\lambda\}_{n \geqq 0}$ *solves* (P_h) *with initial data* $(a_0^\lambda, u_0^\lambda)$, $|\lambda| \leqq \lambda_0$, *and if* a_∞^λ *is any cluster point of* $\{a_n^\lambda\}$, *then*

$$\lim_{\lambda \to 0} a_\infty^\lambda = a_\infty^0, \tag{2.22}$$

provided the whole sequence $\{a_n^0\}$ *converges to* a_∞^0 *and* $A_2(a_\infty^0, \cdot)$ *is uniformly monotone on* C.

Proof. Let $0 < h \leqq h_0$. From (2.2), (2.3) there follows:

$$\left\langle \varepsilon \frac{u_{n+1}^\lambda - u_n^\lambda}{h} + w_{n+1}^\lambda + A_2\left(a_{n+1}^\lambda, u_{n+1}^\lambda\right) - f^*,\ v - u_{n+1}^\lambda \right\rangle + \Psi(v) - \Psi\left(u_{n+1}^\lambda\right) \geqq 0 \tag{2.23}$$

$$\forall v \in C \cap V_N,$$

with some $w_{n+1}^\lambda \in S(u_{n+1}^\lambda)$,

$$\left[\frac{a_{n+1}^\lambda - a_n^\lambda}{h},\ \eta \right] - \left\langle A_2\left(\eta, u_{n+1}^\lambda\right),\ u_{n+1}^\lambda - u^* \right\rangle = 0 \quad \forall \eta \in W_M. \tag{2.24}$$

Let $y_n^\lambda := u_n^\lambda - u_n^0$, $r_n^\lambda := a_n^\lambda - a_n^0$, $q_n^\lambda := u_n^\lambda - u^*$. Then it follows:

$$\left\langle \varepsilon \frac{u_{n+1}^\lambda - u_n^\lambda}{h} + w_{n+1}^\lambda + A_2\left(a_{n+1}^\lambda, u_{n+1}^\lambda\right) - f^*,\ u_{n+1}^\lambda - u_{n+1}^0 \right\rangle + \Psi\left(u_{n+1}^\lambda\right) - \Psi\left(u_{n+1}^0\right)$$

$$\leqq 0 \leqq \left\langle \varepsilon \frac{u_{n+1}^0 - u_n^0}{h} + w_{n+1}^0 + A_2\left(a_{n+1}^0, u_{n+1}^0\right) - f^*,\ u_{n+1}^\lambda - u_{n+1}^0 \right\rangle$$

$$+ \Psi\left(u_{n+1}^\lambda\right) - \Psi\left(u_{n+1}^0\right),$$

whence

$$0 \geqq \left\langle \varepsilon \frac{y_{n+1}^\lambda - y_n^\lambda}{h},\ y_{n+1}^\lambda \right\rangle + \left\langle A_2\left(a_{n+1}^0, y_{n+1}^\lambda\right),\ y_{n+1}^\lambda \right\rangle$$

$$+ \left\langle A_2\left(r_{n+1}^\lambda, u_{n+1}^\lambda\right),\ y_{n+1}^\lambda \right\rangle.$$

Putting $\eta := r_{n+1}^\lambda$ in (2.24) yields:

$$0 = \left[\frac{r_{n+1}^\lambda - r_n^\lambda}{h},\ r_{n+1}^\lambda \right] + \left[\frac{a_{n+1}^0 - a_n^0}{h},\ r_{n+1}^\lambda \right]$$

$$- \left\langle A_2\left(r_{n+1}^\lambda, u_{n+1}^\lambda\right),\ y_{n+1}^\lambda \right\rangle - \left\langle A_2\left(r_{n+1}^\lambda, u_{n+1}^\lambda\right),\ q_{n+1}^0 \right\rangle$$

$$= \left[\frac{r_{n+1}^\lambda - r_n^\lambda}{h},\ r_{n+1}^\lambda \right] - \left\langle A_2\left(r_{n+1}^\lambda, u_{n+1}^\lambda\right),\ y_{n+1}^\lambda \right\rangle - \left\langle A_2\left(r_{n+1}^\lambda, y_{n+1}^\lambda\right),\ q_{n+1}^0 \right\rangle.$$

Addition gives:

$$0 \geqq \left\langle \varepsilon \frac{y_{n+1}^\lambda - y_n^\lambda}{h},\ y_{n+1}^\lambda \right\rangle + \left[\frac{r_{n+1}^\lambda - r_n^\lambda}{h},\ r_{n+1}^\lambda \right] + \left\langle A_2\left(a_{n+1}^0, y_{n+1}^\lambda\right),\ y_{n+1}^\lambda \right\rangle - \left\langle A_2\left(r_{n+1}^\lambda, y_{n+1}^\lambda\right),\ q_{n+1}^0 \right\rangle. \tag{2.25}$$

From (A3ii) we have for every $v \in V$:

$$\left| \left\langle A_2\left(a_\infty^0, v\right),\ v \right\rangle - \left\langle A_2\left(a_{n+1}^0, v\right),\ v \right\rangle \right| \leqq \gamma_1 \left\| a_{n+1}^0 - a_\infty^0 \right\|_{X_0} \|v\|_V^2.$$

Hence there exists a $\hat{\rho} > 0$ and an $n_0 \in \mathbb{N}$ such that:

$$\left\langle A_2\left(a_{n+1}^0, y_{n+1}^\lambda\right),\ y_{n+1}^\lambda \right\rangle \geqq \hat{\rho} \left\| y_{n+1}^\lambda \right\|_V^2 \quad \forall n \geqq n_0. \tag{2.26}$$

Using $\dim W_M < \infty$ we obtain for $n \geqq n_0$:

$$\begin{aligned} &\left\langle \varepsilon \frac{y_{n+1}^\lambda - y_n^\lambda}{h},\ y_{n+1}^\lambda \right\rangle + \left[\frac{r_{n+1}^\lambda - r_n^\lambda}{h},\ r_{n+1}^\lambda \right] \\ &\qquad \leqq -\hat{\rho} \left\| y_{n+1}^\lambda \right\|_V^2 + \gamma_1 \left\| r_{n+1}^\lambda \right\|_{X_0} \left\| y_{n+1}^\lambda \right\|_V \left\| q_{n+1}^0 \right\|_V \\ &\qquad \leqq -\hat{\rho} \left\| y_{n+1}^\lambda \right\|_V^2 + \hat{\rho} \left\| y_{n+1}^\lambda \right\|_V^2 + \frac{1}{4\hat{\rho}} \gamma_1^2 C_1^2 \left\| r_{n+1}^\lambda \right\|_X^2 \left\| q_{n+1}^0 \right\|_V^2 \\ &\qquad \leqq C \left(\frac{\varepsilon}{h} \left\| y_{n+1}^\lambda \right\|_H^2 + \frac{1}{h} \left\| r_{n+1}^\lambda \right\|_X^2 \right) \left\| q_{n+1}^0 \right\|_V^2. \end{aligned}$$

Now we put

$$\left(\phi_{n+1}^\lambda\right)^2 := \frac{\varepsilon}{h} \left\| y_{n+1}^\lambda \right\|_H^2 + \frac{1}{h} \left\| r_{n+1}^\lambda \right\|_X^2.$$

Then

$$\frac{\varepsilon}{h} \left\langle y_n^\lambda,\ y_{n+1}^\lambda \right\rangle + \frac{1}{h} \left[r_n^\lambda, r_{n+1}^\lambda \right] \leqq \phi_n^\lambda \phi_{n+1}^\lambda,$$

and hence,

$$\left(\phi_{n+1}^\lambda\right)^2 \leqq \phi_n^\lambda \phi_{n+1}^\lambda + \gamma_{n+1} \left(\phi_{n+1}^\lambda\right)^2, \quad \text{where } \gamma_{n+1} := C \left\| q_{n+1}^0 \right\|_V^2$$

or

$$\phi_{n+1}^\lambda (1 - \gamma_{n+1}) \leqq \phi_n^\lambda \quad \forall n \geqq n_0.$$

By (2.7) we may assume that $0 \leqq \gamma_{n+1} \leqq \kappa < 1$, $n \geqq n_0$. Hence $\phi_{n+1}^\lambda \leqq (1 + \sigma_{n+1}) \phi_n^\lambda$, $n \geqq n_0$, where

$$\sigma_{n+1} := \frac{\gamma_{n+1}}{1 - \gamma_{n+1}} \leqq \sigma \left\| q_{n+1}^0 \right\|_V^2 \quad \text{for some } \sigma > 0.$$

Induction yields for any $p \in \mathbb{N}$:

$$\phi_{n_0+p}^\lambda \leqq \prod_{k=1}^{p} \left(1 + \sigma_{n_0+k}\right) \phi_{n_0}^\lambda.$$

But

$$\prod_{k=1}^{p}\left(1+\sigma_{n_0+k}\right) \leqq \prod_{k=1}^{p}\exp(\sigma_{n_0+k}) = \exp\left(\sum_{k=1}^{p}\sigma_{n_0+k}\right)$$

$$\leqq \exp\left(\sigma\sum_{k=1}^{\infty}\left\|q_k^0\right\|_V^2\right),$$

which is finite by (2.7). Thus

$$\phi_n^\lambda \leqq \text{const}\cdot\phi_{n_0}^\lambda \quad \text{for all } n\geqq n_0,\ |\lambda|\leqq\lambda_0. \tag{2.27}$$

Next we show that $\lim_{\lambda\to 0}\phi_{n_0}^\lambda = 0$. From (2.25) we have for any $n\leqq n_0$:

$$\left(\phi_{n+1}^\lambda\right)^2 \leqq \phi_n^\lambda\phi_{n+1}^\lambda + \gamma_1\left\|a_{n+1}^0\right\|_{X_0}\left\|y_{n+1}^\lambda\right\|_V^2 + \gamma_1\left\|r_{n+1}^\lambda\right\|_{X_0}\left\|y_{n+1}^\lambda\right\|_V\left\|q_{n+1}^0\right\|_V.$$

By (2.7) the real sequences $\{\|a_{n+1}^0\|_{X_0}\}$ and $\{\|q_{n+1}^0\|_V\}$ are bounded. Thus

$$\left(\phi_{n+1}^\lambda\right)^2 \leqq \phi_n^\lambda\phi_{n+1}^\lambda + C_1\left\|y_{n+1}^\lambda\right\|_H^2 + C_2\left\|r_{n+1}^\lambda\right\|_X^2 \leqq \phi_n^\lambda\phi_{n+1}^\lambda + Ch\left(\phi_{n+1}^\lambda\right)^2.$$

Choosing $h<\hat{h} := 1/C$ yields

$$\phi_{n+1}^\lambda \leqq \frac{1}{1-Ch}\phi_n^\lambda,$$

and thus

$$\phi_{n_0}^\lambda \leqq \left(\frac{1}{1-Ch}\right)^{n_0}\phi_0^\lambda,$$

whence $\lim_{\lambda\to 0}\phi_{n_0}^\lambda = 0$.

The proof is now easily finished: Since a_∞^λ is a cluster point of $\{a_n^\lambda\}$ we have $a_{n_k}^\lambda \to a_\infty^\lambda$ for some subsequence $\{a_{n_k}^\lambda\}$.

Since a_∞^0 is the unique limit point of $\{a_n^0\}$ it follows that $a_{n_k}^0 \to a_\infty^0$, and thus

$$\left\|a_\infty^\lambda - a_\infty^0\right\|_X = \lim_{k\to\infty}\left\|a_{n_k}^\lambda - a_{n_k}^0\right\|_X \leqq C\cdot\phi_{n_0}^\lambda,$$

from which the assertion follows. □

Remark 6. a_∞^0 is uniquely determined under the assumptions of Theorem 2.3 (iii).

Finally we discuss the situation as $\dim V_N\to\infty$. To this end, let $\{V_N\}_{N\geqq 2}$ be a sequence of subspaces of V such that $u_0, u^*\in V_N$, $\forall N\geqq 2$. Assume to each N there is a subspace $W_{M(N)}$ of X_0 of (smallest) dimension $M(N)$ such that $a_0\in W_{M(N)}$ and such that (A7) holds.

Let

$$V_N\subset V_{N+1} \quad \forall N\geqq 2, \quad \bigcup_{N=2}^{\infty} V_N \text{ is dense in } V. \tag{A8}$$

By Theorem 2.3 for any $N\geqq 2$ there exists an $a_\infty^N\in W_{M(N)}$ such that with some $w_\infty^N\in S(u^*)$

$$\left\langle w_\infty^N + A_1\left(a_\infty^N\right) + A_2\left(a_\infty^N, u^*\right) - f^*,\ v-u^*\right\rangle + \Psi(v) - \Psi(u^*) \geqq 0 \quad \forall v\in C\cap V_N. \tag{2.28}$$

The a priori estimate (2.7) gives that $\{a_\infty^N\}$ is bounded in X (recall that the constant C does not depend on V_N and W_M). Thus a subsequence, again denoted by $\{a_\infty^N\}$, converges weakly in X to some $a_\infty \in X$.

Let us assume

(A9) $X \times \{u^*\} \subset D(A_2)$, and $a_n \to a$ weakly in X implies $A_2(a_n, u^*) \to A_2(a, u^*)$ weakly in V^*.

Then $(a_\infty, u^*) \in D(A_2)$. We show that a_∞ is a solution of (P).

By the boundedness of $S(u^*)$ we may assume that $w_\infty^N \to w^*$ weakly in V^* whence $w^* \in S(u^*)$ by the maximal monotonicity. Moreover, from $a_\infty^N \to a_\infty$ weakly in X follows $A_1(a_\infty^N) \to A_1(a_\infty)$ weakly in V^* since every continuous linear mapping is weakly sequentially continuous. We assume further:

(A10) $(V, \|\cdot\|)$ is a Hilbert space, i.e., $\|\cdot\|_V$ is induced by an inner product $(\cdot, \cdot)$ on V.

Now let $v \in C$ be arbitrary. By (A8) there is a sequence $v_N \in V_N$ with $\|v_N - v\|_V \to 0$.

Let P_C denote the $(\cdot, \cdot)$-orthogonal projection operator onto C. As C is closed and convex, P_C is nonexpansive. Thus, $\|v - P_C v_N\|_V \leqq \|v - v_N\|_V \to 0$, as $N \to \infty$, and also, $P_C v_N - u^* \to v - u^*$, strongly in V.

We have

$$\begin{aligned}
&\left\langle w^* + A_1(a_\infty) + A_2(a_\infty, u^*) - f^*,\ v - u^* \right\rangle + \Psi(v) - \Psi(u^*) \\
&\quad = \left[\left\langle w_\infty^N + A_1(a_\infty^N) + A_2(a_\infty^N, u^*) - f^*,\ P_C v_N - u^* \right\rangle + \Psi(P_C v_N) - \Psi(u^*) \right] \\
&\qquad + \left[\left\langle w^* - w_\infty^N + A_1(a_\infty) - A_1(a_\infty^N) + A_2(a_\infty, u^*) - A_2(a_\infty^N, u^*),\ P_C v_N - u^* \right\rangle \right] \\
&\qquad + \Psi(v) - \Psi(P_C v_N) + \left[\left\langle w^* + A_1(a_\infty) + A_2(a_\infty, u^*) - f^*,\ v - P_C v_N \right\rangle \right].
\end{aligned}$$

The first bracket is nonnegative by (2.28), and the second and third brackets approach 0 as $N \to \infty$. Hence passing to the limit as $N \to \infty$ yields

$$\left\langle w^* + A_1(a_\infty) + A_2(a_\infty, u^*) - f^*,\ v - u^* \right\rangle + \Psi(v) - \Psi(u^*) \geqq 0 \quad \forall v \in C,$$

i.e. a_∞ solves (P).

Putting things together we obtain:

THEOREM 2.5. *Let* (A1)–(A10) *hold. If* $\{a_\infty^N\}$ *denotes a sequence of limit points of the solution of* $(P_{h(N)})$, *then each subsequence has a weak cluster point in X which solves* (P).

3. Equations of evolution. We now extend some of the results of the preceding section to equations of evolution type. The notations have the same meaning as in §2. The vector-valued Sobolev spaces $L^2(0, T; H)$, $L^2(0, T; V)$, $L^2(0, T; X)$, $L^2(0, T; V^*)$ and $L^\infty(0, T; X_0)$ together with their norms are defined as usual. Clearly $L^2(0, T; V) \subset L^2(0, T; H) \subset L^2(0, T; V^*)$ with dense and continuous embedding. The dual pairing between elements of $L^2(0, T; V^*)$ and $L^2(0, T; V)$ is denoted by $\langle \cdot | \cdot \rangle$; i.e., for $w \in L^2(0, T; V^*)$ and $v \in L^2(0, T; V)$ we have

$$\langle w | v \rangle = \int_0^T \langle w(t), u(t) \rangle \, dt. \tag{3.1}$$

The inner product $[\cdot | \cdot]$ in the Hilbert space $L^2(0, T; X)$ is defined by

$$[w | v] := \int_0^T [w(t), v(t)] \, dt, \qquad w, v \in L^2(0, T; X). \tag{3.2}$$

Finally we set

$$W^1(0,T;\ V,H) := \Big\{ u \in L^2(0,T;\ V) : u \text{ has a generalized derivative } \frac{du}{dt} \in L^2(0,T;\ V^*) \Big\}. \tag{3.3}$$

As is well known, $W^1(0,T;\ V,H)$ is a Banach space with the norm

$$\|u\|_{W^1} := \|u\|_{L^2(0,T;V)} + \left\| \frac{du}{dt} \right\|_{L^2(0,T;V^*)}. \tag{3.4}$$

Recall that $u \in W^1(0,T;\ V,H)$ implies $u \in C(0,T;\ H)$ (possibly after a modification on a set of zero measure). Now let

$$C := \{ v \in W^1(0,T;\ V,H) : v(0) = \Theta \}. \tag{3.5}$$

An appropriate identification problem for an evolution process is;

(P) Given $u^* \in C$ and $f^* \in L^2(0,T;\ V^*)$, find $a^* \in L^2(0,T;\ X)$ with $(a^*,u^*) \in D(A_2)$ such that there exists $w^* \in S(u^*)$ with

$$\left\langle \frac{du^*}{dt} + w^* + A_1(a^*) + A_2(a^*,u^*) - f^* \,\middle|\, v - u^* \right\rangle + \Psi(v) - \Psi(u^*) \geqq 0 \quad \forall v \in C. \tag{3.6}$$

In correspondence with the assumptions of §2 we assume:

(A1)* $S : D(S) \subset L^2(0,T;\ V) \to 2^{L^2(0,T;V^*)}$ is a maximal monotone graph with $u^* \in \operatorname{int} D(S)$.

(A2)* $A_1 : L^2(0,T;\ X) \to L^2(0,T;\ V^*)$ is linear and bounded.

(A3)* $A_2 : D(A_2) \subset L^2(0,T;\ X) \times L^2(0,T;\ V) \to L^2(0,T;\ V^*)$ is a bilinear map such that

(i) $L^\infty(0,T;\ X_0) \times L^2(0,T;\ V) \subset D(A_2)$,

(ii)

$$\big| \langle A_2(a,u) \,|\, v \rangle \big| \leqq \gamma_1 \|a\|_{L^\infty(0,T;X_0)} \|u\|_{L^2(0,T;V)} \|v\|_{L^2(0,T;V)},$$

$\forall a \in L^\infty(0,T;\ X_0)$, $u, v \in L^2(0,T;\ V)$, for some $\gamma_1 > 0$.

(A4)* $\Psi : L^2(0,T;\ V) \to \mathbb{R}$ is convex and continuous.

Remark 7. The condition $u^*(0) = \Theta$ is merely a normalization.

Remark 8. The operator d/dt with $D(d/dt) := \{ u \in W^1(0,T;\ V,H) : u(0) = \Theta \}$ is maximal monotone from $L^2(0,T;\ V)$ into $L^2(0,T;\ V^*)$.

But as $D(d/dt)$ has empty interior in $L^2(0,T;\ V)$ it appears that the techniques of §2 do not apply unchanged.

Remark 9. If $\|a_n - a\|_{L^\infty(0,T;X_0)} \to 0$ and $\|u_n - u\|_{L^2(0,T;V)} \to 0$, then $A_2(a_n,u_n) \to A_2(a,u)$ weakly in $L^2(0,T;\ V^*)$.

We proceed in close analogy to §2. To this end, let V_N and W_M denote finite-dimensional subspaces of $W^1(0,T;\ V,H)$ and $L^\infty(0,T;\ X_0)$ with

(A5)* $u^* \in V_N$.

(A6)* $\langle A_1(a) + A_2(a,u) \,|\, v \rangle = 0$, $\forall a \in W_M^\perp$, $u, v \in V_N$.

Here $W_M^\perp := \{ z \in L^\infty(0,T;\ X_0) : [z|\eta] = 0, \forall \eta \in W_M \}$.

Condition (A5) is replaced by

(A7)* (P) has at least one solution $a^* \in L^\infty(0,T;\ X_0)$ such that

$$\langle A_2(a^*, u - v) \,|\, u - v \rangle \geqq \delta \|u - v\|^2_{L^2(0,T;V)}$$

$\forall u, v \in L^2(0,T;\ V)$, for some $\delta > 0$.

Now let $\varepsilon>0$ be arbitrary and $h>0$ be fixed. We consider:

(P_h) Given $(a_0,u_0)\in W_M\times(D(S)\cap C\cap V_N)$, find $(a_n,u_n)\in W_M\times(D(S)\cap C\cap V_N)$, $n\in\mathbb{N}$, such that there exists $w_n\in S(u_n)$ with

$$\left\langle \varepsilon\frac{u_{n+1}-u_n}{h}+\frac{d}{dt}u_{n+1}+w_{n+1}+A_1(a_{n+1})+A_2(a_{n+1},u_{n+1})-f^* \mid v-u_{n+1}\right\rangle + \Psi(v)-\Psi(u_{n+1})\geqq 0 \quad \forall v\in C\cap V_N, \tag{3.7}$$

$$\left[\frac{a_{n+1}-a_n}{h}\middle|\eta\right]-\left\langle A_1(\eta)+A_2(\eta,u_{n+1})\mid u_{n+1}-u^*\right\rangle=0 \quad \forall \eta\in W_M. \tag{3.8}$$

With the same technique as in §2 one arrives at the inequality

$$0\geqq\left\langle \varepsilon\frac{q_{n+1}-q_n}{h}\middle|q_{n+1}\right\rangle+\left[\frac{r_{n+1}-r_n}{h}\middle|r_{n+1}\right]+\delta\|q_{n+1}\|^2_{L^2(0,T;V)} +\frac{1}{2}\|q_{n+1}(T)\|^2_H.$$

Here it was used that $q_{n+1}(0)=\Theta$ and thus

$$\left\langle\frac{d}{dt}q_{n+1}\middle|q_{n+1}\right\rangle=\int_0^T\left\langle\frac{d}{dt}q_{n+1}(t),\,q_{n+1}(t)\right\rangle dt =\frac{1}{2}\int_0^T\frac{d}{dt}\|q_{n+1}(t)\|^2_H dt=\frac{1}{2}\|q_{n+1}(T)\|^2_H.$$

As in §2 there follows:

$$\varepsilon\|q_{n+1}\|^2_{L^2(0,T;H)}+\|r_{n+1}\|^2_{L^2(0,T;X)}+2\delta h\|q_{n+1}\|^2_{L^2(0,T;V)} +h\|q_{n+1}(T)\|^2_H\leqq\varepsilon\|q_n\|^2_{L^2(0,T;H)}+\|r_n\|^2_{L^2(0,T;X)}, \tag{3.9}$$

and

$$\sup_{n\in\mathbb{N}}\left\{\varepsilon\|q_n\|^2_{L^2(0,T;H)}+\|r_n\|^2_{L^2(0,T;X)}\right\}+2\delta h\sum_{k=1}^{\infty}\|q_k\|^2_{L^2(0,T;V)} +h\sum_{k=1}^{\infty}\|q_k(T)\|^2_H\leqq C<+\infty, \tag{3.10}$$

where C depends on ε, a_0, u_0, a^*, u^*, but not on h, W_M, V_N.

In analogy to Theorem 2.1 we have:

THEOREM 3.1. *Let* (A1)*–(A7)* *hold. Then there is an* $h_0>0$ *such that* (P_h) *has a solution* $\{(a_n,u_n)\}_{n\geqq 0}$ *for* $0<h\leqq h_0$.

Proof. Let $W_M:=\operatorname{span}\{\Psi_1,\cdots,\Psi_M\}$ with $[\Psi_i|\Psi_j]=\delta_{ij}$, $\forall i,\ j$. Let (a_k,u_k), $0\leqq k\leqq n$, be already constructed. As in the proof of Theorem 2.1 the existence of (a_{n+1},u_{n+1}) is equivalent to the solvability of

$$\left\langle\varepsilon\frac{u-u_n}{h}+w+\frac{du}{dt}+A_1(a)+A_2(a,u)-f^*\mid v-u\right\rangle+\Psi(v)-\Psi(u)\geqq 0,$$

$\forall v\in V_N\cap C$, for some $w\in S(u)$, where

$$a=a_n+h\sum_{k=1}^{M}\left\langle A_1(\Psi_k)+A_2(\Psi_k,u)\mid u-u^*\right\rangle\Psi_k. \tag{3.11}$$

We apply Theorem 2.2 with

$$B := L^2(0,T;\, V), \qquad K := C \cap V_N,$$
$$P_1 := S + \frac{d}{dt}, \qquad g(v) = \Psi(v) - \left\langle \frac{\varepsilon}{h} u_n + f^* \,\middle|\, v \right\rangle,$$

and

$$\begin{aligned} &P_2 : K \to B^*, \\ &P_2(u) := \frac{\varepsilon}{h} u + A_1(a_n) + A_2(a_n, u) \\ &\qquad + h \sum_{k=1}^{M} \left\langle A_1(\Psi_k) + A_2(\Psi_k, u) \,|\, u - u^* \right\rangle \cdot \{ A_1(\Psi_k) + A_2(\Psi_k, u) \}. \end{aligned} \tag{3.12}$$

Clearly $K \neq \varnothing$, and K is convex. Now C is closed in $W^1(0,T;\, V,H)$, and since $K = C \cap V_N$ is contained in the finite-dimensional subspace V_N of $W^1(0,T;\, V,H)$, K is a closed subset of $B = L^2(0,T;\, V)$.

Moreover, S and d/dt are maximal monotone, and $D(d/dt) \cap \operatorname{int} D(S) \ni u^*$. Thus $P_1 = S + d/dt$ is maximal monotone. Finally it is easily checked that P_2 is bounded, continuous, pseudomonotone and coercive with respect to Θ (provided h_0 is sufficiently small, compare the proof of Theorem 2.1). The assertion follows from Theorem 2.2.

THEOREM 3.2. *Let* (A1)*–(A7)* *hold, and let* $0 < h \leqq h_0$. *Let* $\{(a_n, u_n)\}_{n \geqq 0}$ *solve* (P_h). *Then it follows that*:

(i) $\|u_n - u^*\|^2_{L^2(0,T;V)} + \|u_n(T) - u^*(T)\|^2_H \to 0$, *as* $n \to \infty$.

(ii) *Every subsequence of* $\{a_n\}$ *has a cluster point* $a_\infty \in W_M$ *such that with some* $w_\infty \in S(u^*)$

$$\left\langle w_\infty + A_1(a_\infty) + A_2(a_\infty, u^*) + \frac{du^*}{dt} - f^* \,\middle|\, v - u^* \right\rangle + \Psi(v) - \Psi(u^*) \geqq 0 \quad \forall v \in C \cap V_N. \tag{3.13}$$

Proof. (3.10) implies (i). Moreover, any subsequence $\{a_m\}$ has a weak cluster point $a_\infty \in L^2(0,T;\, X)$. Hence we may assume $a_m \to a_\infty$ weakly in $L^2(0,T;\, X)$.

From $\dim W_M < \infty$ follows $\|a_m - a_\infty\|_{L^\infty(0,T;X_0)} \to 0$ and hence $a_\infty \in W_M$. As in the proof of Theorem 2.3 we may assume that $w_m \to w_\infty \in S(u^*)$, weakly in $L^2(0,T;\, V^*)$. From Remark 10 we have

$A_2(a_m, u_m) \to A(a_\infty, u^*)$, weakly in $L^2(0,T;\, V^*)$,

$A_1(a_m) \to A_1(a_\infty)$, weakly in $L^2(0,T;\, V^*)$, by (A2)*.

Relation (3.7) yields for every $v \in C \cap V_N$:

$$\begin{aligned} 0 \leqq \lim_{m \to \infty} \Bigg\{ \Bigg\langle & \varepsilon \frac{u_{m+1} - u_m}{h} + \frac{d}{dt} u_{m+1} + w_{m+1} + A_1(a_{m+1}) \\ & + A_2(a_{m+1}, u_{m+1}) - f^* \,\Big|\, v - u_{m+1} \Bigg\rangle + \Psi(v) - \Psi(u_{m+1}) \Bigg\} \\ = \Big\langle & w_\infty + A_1(a_\infty) + A_2(a_\infty, u^*) - f^* \,\Big|\, v - u^* \Big\rangle + \Psi(v) - \Psi(u^*) \\ & + \lim_{m \to \infty} \left\langle \frac{d}{dt} u_{m+1} \,\middle|\, v - u_{m+1} \right\rangle. \end{aligned}$$

From $v \in W^1(0,T;\, V,H)$ and $u_{m+1} \in C$ there follows:

$$\left\langle \frac{d}{dt}(u_{m+1}-u^*)\middle| v \right\rangle = \int_0^T \left\langle \frac{d}{dt}(u_{m+1}-u^*)(t),\, v(t) \right\rangle dt$$

$$= -\int_0^T \left\langle \frac{d}{dt} v(t),\, u_{m+1}(t)-u^*(t) \right\rangle dt$$

$$+ \left\langle u_{m+1}(T)-u^*(t),\, v(T) \right\rangle \to 0,$$

as $m \to \infty$. Moreover,

$$\left\langle \frac{d}{dt} u_{m+1} \middle| u_{m+1} \right\rangle = \frac{1}{2}\|u_{m+1}(T)\|_H^2 - \frac{1}{2}\|u_{m+1}(0)\|_H^2$$

$$\to \frac{1}{2}\|u^*(T)\|_H^2 - \frac{1}{2}\|u^*(0)\|_H^2 = \left\langle \frac{d}{dt} u^* \middle| u^* \right\rangle,$$

again by (i). Hence (3.13) holds. □

It should be clear that a result which essentially resembles part (iii) of Theorem 2.3 can be proved. Also an analogue of the stability result of Theorem 2.4 is easily established. For the sake of shortness we omit an explicit statement of the result. Instead we turn over our interest to the applications.

4. Applications. In all our applications let $\Omega \subset \mathbb{R}^N$ be open and bounded with a sufficiently smooth boundary $\partial\Omega$. For the sake of shortness we confine ourselves to only some typical examples. There are many other applications in which the developed methods apply as well.

Example 1. Consider the problem:

(4.1) Given $u^* \in \mathring{H}^1(\Omega) \cap H^{1,\infty}(\Omega)$ and $f^* \in H^{-1}(\Omega)$, find a matrix $a^* = (a_{ij}^*)$ with entries $a_{ij}^* \in L^2(\Omega)$ such that: $-\nabla \cdot (a^* \nabla u^*) = f^*$.

This problem was treated extensively in [3], [6]. We make the following specifications: $V = \mathring{H}^1(\Omega)$, $H = L^2(\Omega)$, $V^* = H^{-1}(\Omega)$, $X = \{a = (a_{ij}) : a_{ij} \in L^2(\Omega), \forall i,\, j\}$, $X_0 = \{a \in X : a_{ij} \in L^\infty(\Omega), \forall i,\, j\}$, $C = V$, $\Psi \equiv 0$, $A_1 = \Theta$, $S = \Theta$ and $A_2(a,u) := -\nabla \cdot (a \nabla u)$.

Then (A1), (A2), (A4) are trivial. A_2 is bilinear, and Poincaré's inequality yields (A3ii). (A5) is reasonable from the physical viewpoint. Now, for all $a \in W_M^\perp$:

$$\langle A_2(a,u),\, v \rangle = \int_\Omega \nabla v \cdot (a \nabla u)\, dx = [a, \nabla v \nabla u^T],$$

where the inner product $[\cdot,\cdot]$ of X is given by:

$$[a,b] := \sum_{i,j=1}^N \int_\Omega a_{ij} b_{ij}\, dx.$$

Hence (A7) can be satisfied if only

$$\nabla v \nabla u^T \in W_M \quad \forall u,\, v \in V_N. \tag{4.2}$$

This is possible if $V_N \subset \mathring{H}^1(\Omega) \cap H^{1,\infty}(\Omega)$. Also (A8) can be satisfied since $\mathring{H}^1(\Omega) \cap H^{1,\infty}(\Omega)$ is dense in $\mathring{H}^1(\Omega)$. Finally, from $u^* \in H^{1,\infty}(\Omega)$ it follows that $\nabla v \nabla u^{*T} \in X$ whenever $v \in \mathring{H}^1(\Omega)$. This implies (A9), since from $v \in \mathring{H}^1(\Omega)$ and $a_n \to a$ weakly in X

we get

$$\langle A_2(a_n-a,u^*),\, v\rangle = \int_\Omega \nabla v\cdot\big((a_n-a)\nabla u^*\big)\,dx = \big[a_n-a,\nabla v\nabla u^{*T}\big]\to 0.$$

(A10) is obviously satisfied as well.

Hence the theory of §2 applies. Numerical calculations can be found in [3]. Physical applications are given by stationary states of time-dependent phenomena, such as:

(i) fluid flow through porous media: u = pressure, a = permeability of the medium;
(ii) electricity: u = potential, a = conductivity;
(iii) heat conduction: u = temperature, a = thermal conductivity.

Example 2 (*other boundary conditions*). Let $\sigma^*\in L^\infty(\partial\Omega)$ be nonnegative with $\sigma^*(x)\geqq\sigma>0$ on $\Gamma\subset\partial\Omega$ where meas $(\Gamma)>0$. $\partial/\partial n_a^*$ is the outer conormal derivative at $\partial\Omega$ with respect to the matrix a^*. We consider

(4.3) Given $u^*\in H^{1,\infty}(\Omega)$, $f^*\in L^2(\Omega)$ and $g^*\in H^{-1/2}(\partial\Omega)$, find $a^*\in X$ such that for all $v\in H^1(\Omega)$:

$$\int_\Omega \nabla v\cdot(a^*\nabla u^*)\,dx+\int_\Gamma \sigma^* v u^*\,ds=\int_\Omega f^* v\,dx+\int_{\partial\Omega} g^* v\,ds.$$

Note that from the trace theorems (see [9]) the injection $H^1(\Omega)\to H^{1/2}(\partial\Omega)$ is continuous.

Obviously (4.3) is a weak formulation of

$$(4.4)\qquad -\nabla\cdot(a^*\nabla u^*)=f^* \quad \text{in } \Omega, \qquad \sigma^* u^*+\frac{\partial u^*}{\partial n_a^*}=g^* \quad \text{on } \partial\Omega.$$

We set $V=H^1(\Omega)$, $H=L^2(\Omega)$, $V^*=(H^1(\Omega))^*$, X, X_0 as in Example 1, $A_1=\Theta$, $S=\Theta$, $C=V$, and $\Psi(v):=-\int_{\partial\Omega} g^* v\,ds$, which is convex and continuous by the trace theorems. Next we define A_2 by

$$\langle A_2(a,u),\, v\rangle := \int_\Omega \nabla v\cdot(a\nabla u)\,dx+\int_\Gamma \sigma^* uv\,ds.$$

Since $\sigma^*\geqq\sigma>0$ on a set of positive measure,

$$\|v\|^2 := \int_\Omega |\nabla v|^2 dx+\int_\Gamma \sigma^* |v|^2 ds$$

defines a norm on $H^1(\Omega)$ which is equivalent to $\|\cdot\|_{H^1(\Omega)}$. Hence (A3) follows from

$$\begin{aligned}|\langle A_2(a,u),\, v\rangle| &\leqq C\left(\int_\Omega |\nabla v\cdot\nabla u|dx+\int_\Gamma \sigma^* |uv|\,ds\right)\\ &\leqq C\|u\|\|v\|\leqq \text{const.}\cdot\|u\|_{H^1(\Omega)}\|v\|_{H^1(\Omega)},\end{aligned}$$

if $a\in X_0$, $u,v\in V$.

The other conditions are satisfied as in Example 1. Hence the theory of §2 applies.

Example 3 (*dam problem*). We consider stationary fluid flows through a porous medium $\Omega\subset\mathbb{R}^N$. $\partial\Omega$ is the union of three disjoint sets: Γ^+ (boundary to water reservoirs), Γ^0 (boundary to the atmosphere) and Γ^- (boundary to the impervious ground). See Fig. 1. Let Ω be inhomogeneous and unisotropic, and let Γ^+ be nonempty and relatively open in $\Gamma^+\cup\Gamma^0$.

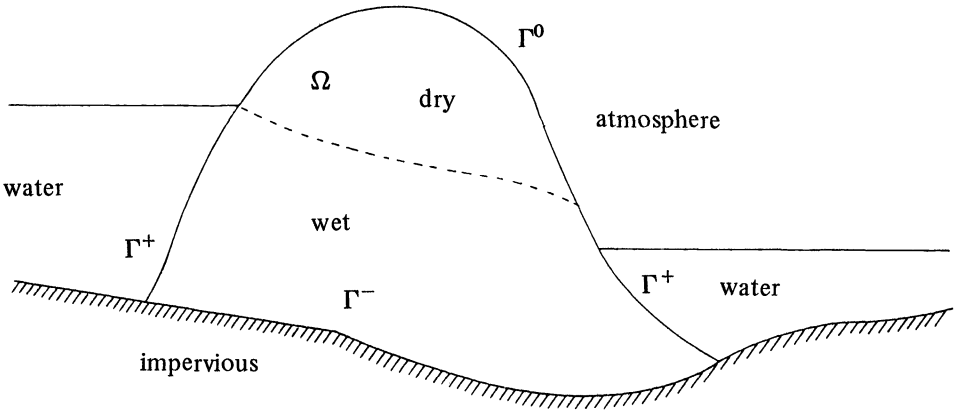


FIG. 1

Assuming the boundary data for pressure u are given by

$$g^* \in H^1(\Omega) \cap H^{1,\infty}(\Omega), \qquad g^* \geqq 0, \tag{4.5}$$

and introducing

$$M(g^*) := \{ v \in H^1(\Omega) : v = g^* \text{ on } \Gamma^+,\ v \leqq g^* \text{ on } \Gamma^0 \},$$

the problem of finding the free boundary between the wet and dry parts of Ω leads to solving:

(4.6) Given $a^* \in X_0$, find $u^* \in M(g^*)$ with $u^* \geqq 0$ and $\gamma^* \in L^\infty(\Omega)$ with $0 \leqq \gamma^* \leqq 1$ and $\gamma^* = 1$ in $\{u^* > 0\}$ such that for all $v \in M(g^*)$:

$$\int_\Omega \nabla(v - u^*) \cdot a^*(\nabla u^* + \gamma^* \vec{e})\, dx \geqq 0.$$

Here X_0, X are as in Example 1, and $\vec{e} = (0, \cdots, 0, 1)$ stands for the gravitational direction. The problem (4.6) has a solution with $u^* \in M(g^*) \cap L^\infty(\Omega)$. For this and a detailed description of the physical background we refer to [1], [2]. We remark that $\{u^* > 0\}$ is the saturated region and that $\{\gamma^* = 0\}$ is the dry part of Ω.

Now set

$$V := \{ v \in H^1(\Omega) : v = 0 \text{ on } \Gamma^+ \}, \qquad K := \{ v \in V : v \leqq 0 \text{ on } \Gamma^0 \}.$$

Then (4.6) can be rewritten as

$$\int_\Omega \nabla(v - u^*) \cdot a^*(\nabla u^* + \nabla g^* + \gamma^* \vec{e})\, dx \geqq 0 \quad \forall v \in K, \tag{4.7}$$

where $u^* \in K$, $u^* \geqq -g^*$. The corresponding parameter identification problem is to reconstruct the permeability matrix:

(4.8) Given $u^* \in K \cap L^\infty(\Omega)$ with $u^* \geqq -g^*$ and $\gamma^* \in L^\infty(\Omega)$ with $0 \leqq \gamma^* \leqq 1$ and $\gamma^* = 1$ in $\{u^* > -g\}$, find $a^* \in X$ such that (4.7) holds.

We show that the theory of §2 applies. $(V, \|\cdot\|_{\mathring{H}^1(\Omega)})$ is a separable and reflexive Banach space with dense and continuous embedding into $H = L^2(\Omega)$. X, X_0 are as in Example 1. Moreover, $f^* = \Theta$, $S = \Theta$, $\Psi \equiv 0$, and $C := K$ is nonempty, closed and convex.

Since $g^* \in H^{1,\infty}(\Omega)$, $\gamma^* \in L^\infty(\Omega)$ the linear mapping A_1 defined by

$$\langle A_1(a), v\rangle := \int_\Omega \nabla v \cdot \big(a\big(\nabla g^* + \gamma^* \vec{e}\big)\big)\, dx \quad \forall v \in V,$$

is bounded. Moreove, the bilinear mapping A_2,

$$\langle A_2(a,u), v\rangle := \int_\Omega \nabla v \cdot (a \nabla u)\, dx \quad \forall v \in V,$$

satisfies (A3). As in Example 1, condition (A7) is fulfilled if only

$$\nabla v\big(\nabla u + \nabla g^* + \gamma^* \vec{e}\big)^T \in W_M \quad \forall u, v \in V_N. \tag{4.9}$$

This is possible if $V_N \subset H^{1,\infty}(\Omega)$ which by (A6) necessitates $u^* \in H^{1,\infty}(\Omega)$.

Example 4 (*linear elasticity*). We follow the lines of [5] and use the notation therein, in particular, the summation convention.

Let $\Omega \subset \mathbb{R}^3$ be the region occupied by an inhomogeneous elastic body with (sufficiently smooth) boundary $\partial\Omega$. Assume Ω is at equilibrium under the influence of forces which are distributed over Ω or act on $\partial\Omega$. We confine ourselves to the static (i.e. time-independent) case. The equations are

$$-\frac{\partial}{\partial x_j}\big(a^*_{ijkh}(x)\varepsilon_{kh}(u)\big) = f_i^* \quad \text{in } \Omega, \qquad i = 1,2,3. \tag{4.10}$$

Here

$$\varepsilon_{ij}(u) := \frac{1}{2}\left(\frac{\partial}{\partial x_i}u_j + \frac{\partial}{\partial x_j}u_i\right), \qquad i = 1,2,3,$$

are the components of the linearized strain tensor in dependence on the displacement vector $u = (u_1, u_2, u_3)$. The vector field $f^* = (f_1^*, f_2^*, f_3^*) \in (L^2(\Omega))^3$ represents a volume density of prescribed forces. $a^*_{ijkh} = a^*_{jihk} = a^*_{khij}$, i, j, k, $h = 1$, 2, 3, are the space-dependent coefficients of elasticity which satisfy the ellipticity condition

(4.11) $a^*_{ijkh}(x)\phi_{ij}\phi_{kh} \geqq \delta \phi_{ij}\phi_{ij}$, a.e. on Ω, for some $\delta > 0$, whenever (ϕ_{ij}) is a (3×3)-matrix.

Typical boundary conditions are (if no friction is present):

$$\partial\Omega = \Gamma_U \cup \Gamma_F, \quad \Gamma_U \cap \Gamma_F = \varnothing, \quad \operatorname{meas}(\Gamma_U) > 0, \tag{4.12}$$

$$u_i = U_i^* \in H^{1/2}(\partial\Omega) \quad \text{on } \Gamma_U, \qquad i = 1,2,3, \tag{4.13}$$

$$\sigma_{ij} n_j = F_i^* \in L^2(\Gamma_F) \quad \text{on } \Gamma_F, \qquad i = 1,2,3. \tag{4.14}$$

Here $n = (n_1, n_2, n_3)$ is the outer unit normal at $\partial\Omega$, $F^* = (F_1^*, F_2^*, F_3^*)$ represents a surface density of forces prescribed on Γ_F, and the components σ_{ij} of the stress tensor are linked to the components of the linearized strain tensor by

$$\sigma_{ij} = a^*_{ijkh}\varepsilon_{kh}(u), \qquad i,j = 1,2,3. \tag{4.15}$$

Now $(H^1(\Omega))^3 = \{v = (v_1, v_2, v_3) : v_i \in H^1(\Omega)\}$ is a Banach space when endowed with the norm

$$\|v\|_{(H^1(\Omega))^3} := \left[\int_\Omega \left\{ v_i v_i + \frac{\partial}{\partial x_j} v_i \frac{\partial}{\partial x_j} v_i \right\} dx\right]^{1/2}$$

By the trace theorem, $U^*=(U_1^*, U_2^*, U_3^*)$ is the trace of a function in $(H^1(\Omega))^3$ which is denoted by U^*, too. The substitution $u=u-U^*$ yields

$$(4.10)^* \qquad -\frac{\partial}{\partial x_j}\left(a_{ijkh}^*(x)\varepsilon_{kh}(u+U^*)\right)=f_i^* \quad \text{in } \Omega, \quad i=1,2,3;$$

$$(4.13)^* \qquad u_i=0 \quad \text{on } \Gamma_U, \qquad i=1,2,3;$$

$$(4.14)^* \qquad a_{ijkh}^*\varepsilon_{kh}(u+U^*)=F_i^*, \qquad i=1,2,3 \quad \text{on } \Gamma_F.$$

Now we define:

$$H := \left(L^2(\Omega)\right)^3,$$
$$V := \left\{v\in\left(H^1(\Omega)\right)^3 : v_i=0, \text{ on } \Gamma_U,\ i=1,2,3\right\},$$
$$X := \left\{a=(a_{ijkh}) : a_{ijkh}\in L^2(\Omega),\ 1\leqq i,\ j,\ k,\ h\leqq 3\right\},$$
$$X_0 := \left\{a\in X : a_{ijkh}\in L^\infty(\Omega),\ 1\leqq i,\ j,\ k,\ h\leqq 3\right\}.$$

X is a Hilbert space when endowed with the inner product

$$(4.16) \qquad [a,b] := \int_\Omega a_{ijkh}b_{ijkh}\,dx.$$

From Korn's inequality it follows (see [5]) that

$$\|v\|_V := \left[\int_\Omega \varepsilon_{ij}(v)\varepsilon_{ij}(v)\,dx\right]^{1/2}$$

defines a norm on V which is equivalent to $\|\cdot\|_{(H^1(\Omega))^3}$.

Now easy calculations using Green's formula (see [5]) show that (4.10)*, (4.12), (4.13)*, (4.14)* is equivalent to:

Given $a^*\in X_0$, $U^*\in(H^1(\Omega))^3$, $f^*\in H$, $F^*\in(L^2(\Gamma_F))^3$, find $u^*\in V$ such that for all $v\in V$:

$$(4.17) \qquad \int_\Omega a_{ijkh}^*\varepsilon_{kh}(u^*+U^*)\varepsilon_{ij}(v)\,dx=\int_\Omega f_i^*v_i\,dx+\int_{\Gamma_F}F_i^*v_i\,ds.$$

Expression (4.17) admits a unique solution ([5, p. 118, Thm. 3.5]). The corresponding identification problem consists in reconstructing the coefficients of elasticity from measurements of the displacements:

(4.18) Given $u^*\in V\cap(H^{1,\infty}(\Omega))^3$, $U^*\in(H^{1,\infty}(\Omega))^3$, $f^*\in H$, $F^*\in(L^2(\Gamma_F))^3$, find $a^*\in X$ with (4.17).

It is easy to check that the assumptions of §2 can be satisfied with the specifications

$$S=\Theta, \quad C=V, \quad \Psi(v) := -\int_{\Gamma_F}F_i^*v_i\,ds,$$

and A_1, A_2 defined by putting for all $v\in V$

$$\langle A_1(a), v\rangle := \int_\Omega a_{ijkh}\varepsilon_{kh}(U^*)\varepsilon_{ij}(v)\,dx,$$

$$\langle A_2(a,u), v\rangle := \int_\Omega a_{ijkh}\varepsilon_{kh}(u)\varepsilon_{ij}(v)\,dx.$$

In this case the assumption (A7) is satisfied if

$$\left(\varepsilon_{ij}(v)\varepsilon_{kh}(u+U^*)\right)\in W_M \quad \forall u,\, v\in V_N.$$

This is possible for $V_N\subset(H^{1,\infty}(\Omega))^3$.

Example 5 (*linear elasticity with friction*). Let in Example 4 the boundary condition (4.14) be replaced by a condition of friction type:

$$(4.14)^{**}\qquad \begin{aligned} &\sigma_N = F_N^* \quad \text{on } \Gamma_F,\\ &|\sigma_T| < \mathbf{F}|F_N^*| \Rightarrow u_T = \Theta,\\ &|\sigma_T| = \mathbf{F}|F_N^*| \Rightarrow \text{there is } \lambda\geqq 0 \text{ such that } u_T = -\lambda\sigma_T. \end{aligned}$$

Let $\operatorname{meas}(\Gamma_F)>0$. $\mathbf{F}\in L^\infty(\Gamma_F)$ is the friction coefficient, and $F_N^*\in L^\infty(\Gamma_F)$ is given. We assume $\mathbf{F}(x)\geqq \mathbf{F}_0>0$, a.e. on Γ_F. The normal stress $\sigma_N=\sigma_{ij}n_jn_i$ is a scalar, and $\sigma_T:=(\sigma_{1T},\sigma_{2T},\sigma_{3T})$ is given by $\sigma_{iT}:=\sigma_{ij}n_j-\sigma_Nn_i$.

Moreover, for $v\in(H^1(\Omega))^3$ we have

$$\begin{aligned} &v_N := v_in_i \text{ (normal displacement, scalar)},\\ &v_T := v-v_Nn \text{ (tangential displacement, vector)}. \end{aligned}$$

Again replacing u by $u-U^*$ one obtains (see [5]) that (4.10), (4.12), (4.13), (4.14)** are (formally) equivalent to:

(4.19) Given $a^*\in X_0$, $f^*\in H$, $U^*\in(H^1(\Omega))^3$, $F_N^*\in L^\infty(\Gamma_F)$ and $\mathbf{F}\in L^\infty(\Gamma_F)$, find $u^*\in V$ such that

$$\int_\Omega a^*_{ijkh}\varepsilon_{kh}(u^*+U^*)\varepsilon_{ij}(v-u^*)\,dx+j(v+U^*)-j(u^*+U^*)$$

$$\geqq \int_\Omega f_i^*(v_i-u_i^*)\,dx+\int_{\Gamma_F}F_N^*(v_N-u_N^*)\,ds \quad \forall v\in V.$$

Here j is the continuous and convex functional

$$(4.20)\qquad j(v):=\int_{\Gamma_F}\mathbf{F}|F_N^*|\,|v_T|\,ds.$$

It should be clear how the corresponding identification problem is formulated and how it fits into the framework of §2.

Remark 10. We have not given any examples for equations of evolution. In [7] the simplest case of identifying the system matrix of a system of ordinary differential equations is treated, and numerical results are given.

It should be clear from the preceding examples how for example the problem of reconstructing a time-dependent thermal conductivity matrix in a linear heat conduction problem can be fitted into the setting of §3.

REFERENCES

[1] H. W. ALT, *Strömungen durch inhomogene poröse Medien mit freiem Rand*, J. Reine Angew. Math., 305 (1979), pp. 89–115.

[2] ———, *Numerical solution of steady-state porous flow free boundary problems*, Numer. Math., 36 (1980), pp. 73–98.

[3] H. W. ALT, K.-H. HOFFMANN AND J. SPREKELS, *A numerical procedure to solve certain identification problems*, Intern. Ser. Numer. Math., 68 (1984), pp. 11–43.

[4] W. H. CHEN AND J. H. SEINFELD, *Estimation of spatially varying parameters in partial differential equations*, Int. J. Control, 15 (1972), pp. 487–495.

[5] G. DUVAUT AND J. L. LIONS, *Inequalities in Mechanics and Physics*, Springer-Verlag, Berlin-Heidelberg-New York, 1976.

[6] K.-H. HOFFMANN AND J. SPREKELS, *On the identification of elliptic problems by asymptotic regularization*, Numer. Funct. Anal. Optim., 7 (1984/85), pp. 157–178.

[7] ———, *Towards the identification of ordinary differential equations from measurements*, Control and Cybernetics, to appear.

[8] R. KLUGE, *Nichlineare Variationsungleichungen und Extremalaufgaben*, VEB Deutscher Verlag der Wissenschaften, Berlin, 1979.

[9] J. L. LIONS AND E. MAGENES, *Non-homogeneous Boundary Value Problems and Applications, Vol.* I, Springer Verlag, Berlin-Heidelberg-New York, 1972.

SIAM J. MATH. ANAL.
Vol. 17, No. 5, September 1986

011

REAL CONTINUED FRACTIONS AND ASYMPTOTIC EXPANSIONS*

BURNETT MEYER†

Abstract. Let $\mathbf{K}(a_n(x)/b_n(x))$ be a continued fraction, where $a_n(x)$ and $b_n(x)$ are polynomials with nonnegative coefficients, in a real variable x. Let the continued fraction correspond at $x=0$ to a formal power series in x or at $x=\infty$ to a formal power series in x^{-1}. Conditions are given which insure that the corresponding series are asymptotic expansions of the functions to which the odd and even parts of the continued fraction converge as $x\to 0+$ or as $x\to+\infty$.

Key words. continued fraction, asymptotic expansion, correspondence of power series

AMS(MOS) subject classifications. Primary 30B70, 30E15, 40A15

In this paper the author continues his study of a question posed by Jones and Thron [2, p.331]: Let a formal power series $\sum c_k z^{-k}$ be given, and let $\mathbf{K}(a_n(z)/b_n(z))$ be a continued fraction that corresponds to the above series at $z=\infty$ and that converges to a holomorphic function $f(z)$ in a region D with $z=\infty$ on its boundary. Is the given series the asymptotic expansion of $f(z)$ at $z=\infty$, with respect to D? A similar question can be asked if the correspondence is at $z=0$.

1. Preliminaries. We use the notation $\mathbf{K}_{n=1}^{\infty}(a_n/b_n)$ for the continued fraction

$$\frac{a_1}{b_1}+\frac{a_2}{b_2}+\cdots+\frac{a_n}{b_n}+\cdots.$$

In the following two definitions, asymptotic expansions and correspondence will be defined at a finite point z_0, with the modifications necessary for $z_0=\infty$ put in parentheses.

DEFINITION [1, p. 359]. Let f be a function defined on a set S in the complex plane with $z=z_0$ $(z=\infty)$ a limit point of S. Let $L=\sum_{k=0}^{\infty}c_k(z-z_0)^k$ $(L=\sum_{k=0}^{\infty}c_k z^{-k})$ be a formal power series in $z-z_0$ (z^{-1}). Let $G_n(z)=\sum_{k=0}^{n}c_k(z-z_0)^k$ $(G_n(z)=\sum_{k=0}^{n}c_k z^{-k})$. Then L is the asymptotic expansion of f as $z\to z_0$ $(z\to\infty)$, $z\in S$, if $f(z)-G_n(z)=O((z-z_0)^{n+1})$ $(f(z)-G_n(z)=O(z^{-n-1}))$, as $z\to z_0$ $(z\to\infty)$, $z\in S$, for $n=0,1,2,\cdots$.

DEFINITION [2, p. 149]. Let $\{f_n\}$ be a sequence of complex-valued functions of a complex variable z, each holomorphic at $z=z_0$ $(z=\infty)$. Let $L=\sum_{k=0}^{\infty}c_k(z-z_0)^k$ $(L=\sum_{k=0}^{\infty}c_k z^{-k})$ be a formal power series in $z-z_0$ (z^{-1}), and let

$$G_m(z)=\sum_{k=0}^{m}c_k(z-z_0)^k\qquad\left(G_m(z)=\sum_{k=0}^{m}c_k z^{-k}\right).$$

The sequence $\{f_n\}$ is said to *correspond* to L at $z=z_0$ $(z=\infty)$, with order of correspondence ν_n, if there exists a sequence $\{\nu_n\}$ of positive integers such that $\nu_n\to\infty$ and

$$f_n(z)-G_{\nu_n-1}(z)=O\big((z-z_0)^{\nu_n}\big)\qquad\big(f_n(z)-G_{\nu_n-1}(z)=O(z^{-\nu_n})\big)$$

as $z=z_0$ $(z\to\infty)$.

*Received by the editors April 15, 1985.
†Department of Mathematics, University of Colorado, Boulder, Colorado 80309.

A continued fraction is said to correspond to a formal power series if the sequence of its approximants corresponds to the series.

THEOREM 1 [4, p. 48]. *If $a_n>0$ and $b_n>0$ for all n, then the even approximants of $\mathbf{K}(a_n/b_n)$ form an increasing sequence, with limit k. The odd approximants form a decreasing sequence with limit K, and $0<k\leqq K<\infty$.*

The following lemma is probably not new, but the author is unable to find a reference.

LEMMA 1. *Let $\sum_{k=0}^{\infty}c_k(z-z_0)^k$ be a formal power series and let $G_m(z)=\sum_{k=0}^{m}c_k(z-z_0)^k$. If $f(z)-G_m(z)=O((z-z_0)^{m+1})$ as $z\to z_0$, $z\in S$ holds for $m=m_1$, then it holds for all nonnegative $m\leqq m_1$.*

Proof.

$$\left|f(z)-G_{m_1}(z)\right|=\left|f(z)-G_{m_1-1}(z)-c_{m_1}(z-z_0)^{m_1}\right|.$$

So there is a constant γ_{m_1+1} such that

$$\left|f(z)-G_{m_1-1}(z)\right|-\left|c_{m_1}(z-z_0)^{m_1}\right|\leqq\gamma_{m_1+1}\left|z-z_0\right|^{m_1+1}.$$

Hence, $f(z)-G_{m_1-1}(z)=O((z-z_0)^{m_1})$ as $z\to z_0$, $z\in S$. The proof is completed by repetition of the argument. □

The analogous lemma for $z_0=\infty$ is similarly proved.

2. The following theorem is an improvement on [3, Thm. 1].

THEOREM 2. *Let $\{f_n\}$ be a sequence of functions, holomorphic at z_0 and meromorphic in a domain D, with $z_0\in D$. Let z_0 be a limit point of a set $S\subset D$. Let $\{f_n\}$ correspond at z_0 to a formal power series $L=\sum_{k=0}^{\infty}c_k(z-z_0)^k$, and let $\nu_1\leqq\nu_2\leqq\cdots\leqq\nu_n\leqq\cdots$, with $\nu_n\to\infty$, where ν_n is the order of correspondence. If there exist a function f, defined on S, and positive constants k_n $(n=0,1,2,\cdots)$ such that*

$$\left|f(z)-f_n(z)\right|\leqq k_n\left|f_{n+1}(z)-f_n(z)\right| \tag{1}$$

for $z\in S$ and $n=0,1,2,\cdots$, then L is the asymptotic expansion of f as $z\to z_0$, $z\in S$.

Proof. Let $G_n(z)=\sum_{k=0}^{n}c_k(z-z_0)^k$. Then

$$\begin{aligned}\left|f(z)-G_n(z)\right|&\leqq\left|f(z)-f_n(z)\right|+\left|f_n(z)-G_n(z)\right|\\&\leqq k_n\left|f_{n+1}(z)-f_n(z)\right|+\left|\beta(z-z_0)^{\mu_n}+\cdots\right|,\end{aligned}$$

where $\mu_n=\min(\nu_n,n+1)$. Thus, $f(z)-G_n(z)=O((z-z_0)^{\mu_n})$ as $z\to z_0$, $z\in S$, for $n=0,1,2,\cdots$.

So

$$\begin{aligned}&\left|f(z)-G_{\mu_n-1}(z)-c_{\mu_n}(z-z_0)^{\mu_n}-\cdots-c_n(z-z_0)^n\right|\\&\qquad\leqq k_n\left|\alpha(z-z_0)^{\mu_n}+\cdots\right|+\left|\beta(z-z_0)^{\mu_n}+\cdots\right|.\end{aligned}$$

Thus, $f(z)-G_{\mu_n-1}(z)=O((z-z_0)^{\mu_n})$ as $z\to z_0$, $z\in S$, for $n=0,1,2,\cdots$. Since $\mu_n\to\infty$, we apply Lemma 1 to obtain

$$f(z)-G_{m-1}(z)=O\big((z-z_0)^m\big)\quad\text{as } z\to z_0,\ z\in S,\text{ for } m=0,1,2,\cdots.$$

(In the above proof we assumed $\nu_n<n+1$. If $\nu_n\geqq n+1$, the proof is easier.) □

A similar theorem can be proved for the case $z_0=\infty$, considering expansions of the form $\sum_{k=0}^{\infty} c_k z^{-k}$.

Also, the right side of (1) may be replaced by $k_n|f_n(z)-f_{n-1}(z)|$. See [3].

3. Real continued fractions. We now consider the real continued fraction $\mathbf{K}(a_n(x)/b_n(x))$, where a_n and b_n are polynomials in a real variable x. We will obtain simple "rules of thumb" for concluding that the corresponding power series at 0 or at ∞ are the asymptotic expansions of the limits of the continued fraction as $x\to 0+$ or as $x\to+\infty$, respectively.

Given the polynomial $P(x)=a_n x^n+\cdots+a_0$, $a_n\neq 0$, $n\geqq 0$. Let $d(P)=n$, the degree of P. Denote the degree of the nonzero term of P of *lowest* degree by $l(P)$. We call a polynomial nonnegative if all its coefficients are nonnegative.

THEOREM 3. *Let $a_n(x)$ and $b_n(x)$ be nonnegative polynomials. If $l(b_n(x))=0$ for all n and if $\sum_{k=1}^{n} l(a_k(x))\to\infty$ as $n\to\infty$, then the power series corresponding to $\mathbf{K}(a_n(x)/b_n(x))$ at $x=0$ is the asymptotic expansion of the limits of the odd and even parts of the continued fraction as $x\to 0+$.*

Proof. Let the nth approximant of the continued fraction be denoted by $f_n(x)=A_n(x)/B_n(x)$. Since $l(b_n)=0$, it is easily shown by induction, using the recursion formula for B_n [2, p. 20], that the constant term of $B_n(x)$ is >0. Now, by the determinant formula [2, p. 20]

$$\frac{A_{n+1}}{B_{n+1}}-\frac{A_n}{B_n}=\frac{(-1)^n\prod_{k=1}^{n+1}a_k(x)}{B_n(x)B_{n+1}(x)}=\frac{P(x)}{Q(x)}, \tag{2}$$

where $l(P(x))=\sum_{k=1}^{n+1} l(a_k(x))$ and $Q(x)$ is a polynomial with constant term >0. Therefore, the continued fraction corresponds to a power series, with order of correspondence $\sum_{k=1}^{n+1} l_k$.

By Theorem 1, the odd and even parts of the continued fraction converge to functions g_1 and g_2, respectively. The following inequalities hold:

$$f_{2n}(x)\leqq g_2(x)\leqq g_1(x)\leqq f_{2n+1}(x)$$

for $n=1,2,\cdots$ and $x>0$. Thus,

$$|g_k(x)-f_n(x)|\leqq|f_{n+1}(x)-f_n(x)|$$

for $k=1,2$, $n=1,2,\cdots$, and $x>0$. By Theorem 2 the corresponding series is the asymptotic expansion of g_1 and of g_2 as $x\to 0+$. □

We note that Theorem 3 applies to all positive C-fractions (not just the *regular* positive C-fractions), SITZ fractions, g-fractions, associated continued fractions for which the k_n are negative and the l_n are positive, and positive T-fractions. See [2, pp. 386–394] for definitions of these kinds of continued fractions.

THEOREM 4. *Let $a_n(x)$ and $b_n(x)$ be nonnegative polynomials of degrees d_n and e_n, respectively. Then the power series in z^{-1} corresponding to $\mathbf{K}(a_n(x)/b_n(x))$ at $x=\infty$ is the asymptotic expansion of the continued fraction as $x\to+\infty$, provided*

$$h_n=2\sum_{k=1}^{n} e_k+e_{n+1}-\sum_{k=1}^{n+1} d_k\to\infty \tag{3}$$

as $n\to\infty$. If $d_n\leqq e_n$ for all n, condition (3) may be replaced by the simpler condition $\sum_{k=1}^{n+1} e_k\to\infty$ as $n\to\infty$.

Proof. We use the notation of Theorem 3. Using induction and the recursion formula for B_n, we can prove that $d(B_n) \geqq e_1 + e_2 + \cdots + e_n$. We again use formula (2). We have $d(P) = \sum_{k=1}^{n+1} d_k$ and $d(Q) \geqq 2\sum_{k=1}^{n} e_k + e_{n+1}$. Thus, the continued fraction corresponds to a power series in z^{-1}, with order of correspondence $\geqq h_n$. The remainder of the proof is similar to the proof of Theorem 3. □

Theorem 4 applies to modified S-fractions, H-fractions if all the b_n are positive, J-fractions if all the c_n^2 are negative and all the d_n are positive, and positive T-fractions. And, of course, Theorems 3 and 4 apply to many continued fractions that have not been classified and given names.

4. An example. The continued fraction

$$\frac{1}{1} + \frac{x}{1} + \frac{x}{1} + \frac{2x}{1} + \frac{2x}{1} + \frac{3x}{1} + \cdots$$

corresponds at $x = 0$ to the series $\sum_{n=0}^{\infty} (-1)^n n! x^n$. [1, p. 519] By Theorem 4.58 in [2, p. 136], it follows that the continued fraction converges for all positive x. So, by Theorem 3, the series is the asymptotic expansion of the limit of the continued fraction, which is

$$\int_0^{\infty} \frac{e^{-s}\, ds}{1 + xs},$$

as $x \to 0+$.

REFERENCES

[1] P. Henrici, *Applied and Computational Complex Analysis*, vol. 2, John Wiley, New York, 1977.

[2] W. B. Jones and W. J. Thron, *Continued Fractions: Analytic Theory and Applications*, Encyclopedia of Mathematics and its Applications, vol. 11, Addison-Wesley, Reading, MA, 1980.

[3] B. Meyer, *On continued fractions corresponding to asymptotic series*, Rocky Mountain J. Math., 15 (1985), pp. 167–172.

[4] O. Perron, *Die Lehre von Kettenbrüchen*, vol. 2, Teubner, Stuttgart, 1957.

SIAM J. MATH. ANAL.
Vol. 17, No. 5, September 1986

012

BOREL SUMMABILITY AND CONVERGING FACTORS FOR SOME EVERYWHERE DIVERGENT SERIES*

AVRAM SIDI†

Abstract. In this work we deal with the problem of interpretation of certain classes of everywhere divergent power series within the framework of Borel summability, and derive asymptotic expansions for their partial sums and/or their converging factors when the number of terms in the partial sums goes to infinity.

Key words. divergent series, Borel summability, converging factors, asymptotic expansion

AMS(MOS) subject classifications. Primary 34E05, 40A05, 40D15, 40G10, 65B05, 65B10

1. Introduction. Consider the formal power series $F(\zeta):=\sum_{r=1}^{\infty}a_r\zeta^r$ whose terms are of the form

$$a_r=r^p w(r)(r!)^m, \tag{1.1}$$

where $p\geqq 0$ and $m\geqq 1$ are integers, and $w(r)$ is such that

$$w(r)\sim\sum_{i=0}^{\infty}\frac{w_i}{r^{i+\sigma}}\quad\text{as } r\to\infty,\quad\text{for some } \sigma>0, \tag{1.2}$$

with w_i being some constants independent of r. Obviously $F(\zeta)$ does not converge for any value of ζ. In this work we shall be concerned with the interpretation of the "sum" of $F(\zeta)$, and with the asymptotics of the converging factor of the partial sum $A_n(\zeta)=\sum_{i=1}^{n}a_i\zeta^i$ in the limit $n\to\infty$, for ζ fixed. This problem arises when one tries to apply the t or u transformation of Levin [4] to the sequence $A_j(\zeta)$, $j=1,2,\cdots$, to obtain an approximation to the "sum" of $F(\zeta)$, or to the anti-limit of the sequence $\{A_j(\zeta)\}$.

The t and u transformations are nonlinear methods for accelerating the convergence of a slowly converging sequence to its limit, or for effecting convergence of a diverging sequence to its anti-limit. There is ample numerical evidence (see the numerical examples given in [10]) that suggests that in order for the t (or u) transformation to be efficient on a sequence B_i, $i=1,2,\cdots$, B_i has to be of the form

$$B_{r-1}=B+R_r f(r), \tag{1.3}$$

where B is the limit or anti-limit of $\{B_i\}$, and $f(r)$ should be such that

$$f(r)\sim\sum_{i=0}^{\infty}\frac{\beta_i}{r^i}\quad\text{as } r\to\infty \tag{1.4}$$

and $R_r=r^{\lambda}b_r$, where $b_1=B_1$, $b_r=B_r-B_{r-1}$, $r\geqq 2$, and $\lambda=0$ for the t transformation (or $\lambda=1$ for the u transformation). From the conjectured behavior of B_i in (1.3) and (1.4), it follows that

$$b_{r+1}=c(r)b_r, \tag{1.5}$$

* Received by the editors October 19, 1982, and in revised form February 21, 1985.
† Computer Science Department,Technion–Israel Institute of Technolgy, Haifa, Israel.

where

$$c(r)=\frac{1+r^{\lambda}f(r)}{(r+1)^{\lambda}f(r+1)}\sim\sum_{i=0}^{\infty}\frac{c_i}{r^{i-q}}\quad\text{as } r\to\infty, \tag{1.6}$$

with q being an integer. The solution of (1.5) and (1.6) is known to be, see [3, p. 70], of the form $b_r=x^r v(r)(r!)^q$, with $v(r)$ being such that $v(r)\sim\sum_{i=0}^{\infty}v_i/r^{i+\alpha}$ as $r\to\infty$, for some α, cf. (1.1) and (1.2). With the help of [7, Thm. 6.1], it has been proved in [8, Thm. 2.2] that when $q=0$, and $\lim_{i\to\infty}B_i$ exists, i.e. $|x|>1$, the B_i satisfy (1.3) and (1.4). When $\lim_{i\to\infty}B_i$ does not exist, i.e., $|x|>1$ or $|x|\geqq 1$, it is not known, in general, whether (1.3) and (1.4) still hold, although, under certain circumstances, the techniques of the present work can be used to show that they do. This will be indicated at the end of §2. Using the technique of the proof of [8, Thm. 2.2], (1.3) and (1.4) can be shown to hold for all integers $q\leqq -1$ and for all x, since for this case $\lim_{i\to\infty}B_i$ exists for all x. For $q>0$, in which case $\lim_{i\to\infty}B_i$ does not exist for any x, no result like (1.3) and (1.4) is known in general, and precisely this is the subject of the present work. For $q=1$, (1.3) and (1.4) have been shown to hold for two special cases, see [9].

In the present work we actually show that under certain conditions, $A_i(\zeta)$ is of the form

$$A_{r-1}(\zeta)=A(\zeta)+a_r\zeta^r g(r,\zeta), \tag{1.7}$$

where $A(\zeta)$ is the Borel-type sum of $F(\zeta)$ (to be defined later), and the converging factor $g(r,\zeta)$ has an asymptotic expansion of the form

$$g(r,\zeta)\sim\sum_{i=0}^{\infty}\frac{g_i(\zeta)}{r^i}\quad\text{as } r\to\infty, \tag{1.8}$$

with $g_i(\zeta)$ being polynomials in ζ^{-1}. We also note that all of the numerical examples of everywhere divergent series considered in [11, Tables A3 and A4] are of the form above with $q>0$, and for these examples Levin's transformations produce accurate approximations to their Borel-type sums.

A similar but less general approach to the interpretation of divergent series has been introduced by Dingle in a series of papers, and this approach is summarized in his book [2, Chaps. XXI and XXII]. Dingle is concerned with summing the remainder series $\sum_{i=r}^{\infty}a_i\zeta^i$ for fixed r, whereas our main concern is with the asymptotics of it as $r\to\infty$. Olver's book [6, Chap. 14] contains another approach to the estimation of $\sum_{i=r}^{\infty}a_i\zeta^i$ that was introduced by Stieltjes, and developed further by Airy and J. C. P. Miller; see [6] for further references. Again our problem is different than that considered in Olver's book.

2. Theory.

LEMMA 2.1. *Define*

$$Q_k(r,z)=\left(z\frac{d}{dz}\right)^k\frac{z^r}{1-z}. \tag{2.1}$$

Then

$$Q_k(r,z)=\frac{\sum_{i=0}^{k}g_{k,i}(r)z^{r+i}}{(1-z)^{k+1}}, \tag{2.2}$$

where $g_{k,i}(r)$ are polynomials of degree k in r, satisfying

$$(2.3)\qquad \begin{aligned} &g_{k,0}(r)=r^k, \qquad g_{k,k}(r)=(1-r)^k,\\ &g_{k,i}(r)=(r+i)g_{k-1,i}(r)+(k-r-i+1)g_{k-1,i-1}(r), \qquad i=1,\cdots,k-1. \end{aligned}$$

Proof. Equation (2.3) follows easily by induction on k, starting with $k=0$ and $g_{0,0}(r)=1$. □

THEOREM 2.2. *Let a_r be expressible in the form*

$$(2.4)\qquad a_r=r^p w(r)\prod_{i=1}^{m}(\mu_i r+\nu_i)!,$$

where we assume that $p\geqq 0$ is an integer,

$$(2.4a)\qquad w(r)=\int_0^\infty e^{-rt}\varphi(t)\,dt, \qquad r\geqq 1,$$

for some function $\varphi(t)$ such that $\int_0^\infty e^{-t}|\varphi(t)|\,dt<\infty$, and μ_i and ν_i satisfy

$$(2.4b)\qquad \mu_i>0, \quad \mu_i+\nu_i>-1, \quad i=1,\cdots,m.$$

Obviously the power series $F(\zeta):=\sum_{r=1}^\infty a_r\zeta^r$ diverges for all $\zeta\neq 0$. For $0<\theta_0<\pi$ define the bounded sectors $S(\rho,\theta_0)$ in the ζ-plane by

$$(2.5)\qquad S(\rho,\theta_0)=\left\{\zeta=|\zeta|e^{i\theta}\colon |\zeta|\leqq\rho,\ \theta_0\leqq\theta\leqq 2\pi-\theta_0\right\}.$$

Then $F(\zeta)$ is the asymptotic expansion of its Borel-type sum

$$(2.6)\qquad \mathcal{F}(\zeta)=\int_0^\infty\cdots\int_0^\infty \psi(\vec{t})\left(z\frac{d}{dz}\right)^p\left(\frac{z}{1-z}\right)\prod_{i=0}^{m}dt_i,$$

as $\zeta\to 0$, uniformly in $S(\rho,\theta_0)$, for each finite ρ, where $\vec{t}=(t_0,t_1,\cdots,t_m)$, and

$$(2.6a)\qquad \psi(\vec{t})=\exp\left(-\sum_{i=1}^{m}t_i\right)\left(\prod_{i=1}^{m}t_i^{\nu_i}\right)\varphi(t_0),$$

$$(2.6b)\qquad z=\zeta e^{-t_0}\prod_{i=1}^{m}t_i^{\mu_i}.$$

The function $\mathcal{F}(\zeta)$ is analytic in the ζ-plane cut along $[0,\infty)$.

Remarks. (1) If $\varphi(t)\sim\sum_{i=0}^\infty\varphi_i t^{i+\sigma-1}$ as $t\to 0+$, with $\sigma>0$, the application of Watson's lemma, see [6, p. 71], yields $w(r)\sim\sum_{i=0}^\infty\varphi_i(i+\sigma-1)!/r^{i+\sigma}$ as $r\to\infty$, and this is exactly of the form given in (1.2) with $w_i=\varphi_i(i+\sigma-1)!$, $i=0,1,\cdots$. Furthermore, if we take $\mu_i=1$, $\nu_i=0$, $i=1,\cdots,m$, then we are back at (1.1) and (1.2).

(2) There is no loss of generality in assuming $\mu_i+\nu_i>-1$ in (2.4b). For, if $\mu_i+\nu_i>-1$ is not satisfied for all i, we can consider the series $F'(\zeta):=\sum_{r=1}^\infty a'_r\zeta^r$, where $a'_r=a_{r+k}$, with k being chosen such that $\mu_i+(k\mu_i+\nu_i)>-1$, $1\leqq i\leqq m$. Note that $F(\zeta)=A_k(\zeta)+\zeta^k F'(\zeta)$.

(3) The Borel-type sum $\mathcal{F}(\zeta)$ given in (2.6) is obtained by substituting (2.9) (see proof below) in $\sum_{r=1}^\infty a_r\zeta^r$, interchanging the summation with all the integrations, and then summing the geometric-type series $M(z)=\sum_{r=1}^\infty r^p z^r$ to obtain $M(z)=(zd/dz)^p(z/(1-z))$. It can be shown that the Borel sum of $F(\zeta)$, namely $\int_0^\infty e^{-t}(\sum_{r=1}^\infty a_r t^r\zeta^r/r!)\,dt$, see [3, p. 78], is its Borel-type sum when $m=1$, $\mu_1=1$, and $\nu_1=0$.

(4) If in (2.4) $r^p w(r)=Cr^{p'}$, where C is a constant and $p'=p-1\geqq 0$, then the Borel-type sum in (2.6) reduces to that obtained from (2.6) by omitting the integration with respect to t_0 after $\varphi(t_0)$ in $\psi(\vec{t})$ has been replaced by C, p has been replaced by p', and the factor e^{-t_0} has been deleted from z. This can be shown by observing that actually $w(r)=C/r$, thus $\varphi(t_0)=C$, and performing the integral with respect to t_0, reducing (2.6) to an m-dimensional integral.

Proof. Using the fact that

$$\frac{z}{1-z}=z+z^2+\cdots+z^{r-1}+\frac{z^r}{1-z}, \tag{2.7}$$

we have

$$\left(z\frac{d}{dz}\right)^p\frac{z}{1-z}=\sum_{j=1}^{r-1}j^p z^j+Q_p(r,z). \tag{2.8}$$

Letting z be as in (2.6b), and substituting (2.8) in (2.6), and using the fact that

$$j^p\int_0^\infty\cdots\int_0^\infty\psi(\vec{t})z^j\prod_{i=0}^m dt_i=a_j\zeta^j,\qquad j=1,2,\cdots, \tag{2.9}$$

we obtain

$$\mathscr{F}(\zeta)=A_{r-1}(\zeta)+U_r(\zeta), \tag{2.10}$$

where

$$U_r(\zeta)=\int_0^\infty\cdots\int_0^\infty\psi(\vec{t})Q_p(r,z)\prod_{i=0}^m dt_i. \tag{2.11}$$

From Lemma 2.1

$$Q_p(r,z)=\frac{\sum_{i=0}^p g_{p,i}(r)z^{r+i}}{(1-z)^{p+1}}. \tag{2.12}$$

It is easy to see that

$$|1-z|\geqq\sin\theta_0,\quad \text{all } \zeta\in S(\rho,\theta_0),\quad \text{all } \rho>0. \tag{2.13}$$

Substituting (2.12) in (2.11), taking the modulus of both sides, and using (2.13), we obtain

$$\begin{aligned}|U_r(\zeta)|\leqq(\sin\theta_0)^{-p-1}&\sum_{i=0}^p|g_{p,i}(r)||\zeta^{r+i}|\\&\times\left(\prod_{j=1}^m[\mu_j(r+i)+\nu_j]!\right)\int_0^\infty e^{-(r+i)t}|\varphi(t)|\,dt,\end{aligned} \tag{2.14}$$

which, for all $\zeta\in S(\rho,\theta_0)$, becomes

$$|U_r(\zeta)|\leqq K_r|\zeta|^r, \tag{2.15}$$

with K_r being *independent* of ζ. This proves the first part of the theorem. The second part of the theorem is obvious. □

We now go on to analyze the "remainder" term $U_r(\zeta)$ in the limit $r\to\infty$.

THEOREM 2.3. *Assume that all the conditions of Theorem* 2.2 *are satisfied, and, in addition,* $\varphi(t)$ *is continuous in a neighborhood of* 0 *except possibly at* 0, *and satisfies*

$$\varphi(t)\sim\varphi_0 t^{\sigma-1} \quad \text{as } t\to 0^+, \quad \text{for some } \sigma>0. \tag{2.16}$$

Then, for any integer $k\geqq 0$,

$$U_r(\zeta)=-\sum_{j=0}^{k-1} a_{r-1-j}\zeta^{r-1-j}+0\left(a_{r-1-k}\zeta^{r-1-k-p}\right) \quad \text{as } r\to\infty, \tag{2.17}$$

uniformly in ζ *for* $\zeta\in S(\rho,\theta_0)$, *for each finite* ρ.

Proof. Expressing $Q_p(r,z)$ (see (2.1)) in the form

$$Q_p(r,z)=-\left(z\frac{d}{dz}\right)^p\frac{z^{r-1}}{1-1/z}, \tag{2.18}$$

and making use of (2.7) with z replaced by $1/z$, we have, for any integers $k\geqq 1$ and $N\geqq k$,

$$Q_p(r,z)=-\sum_{j=0}^{N-1}(r-1-j)^p z^{r-1-j}+Q_p(r-N,z). \tag{2.19}$$

Substituting (2.19) in (2.11), and invoking (2.9), we obtain

$$U_r(\zeta)=-\sum_{j=0}^{N-1} a_{r-1-j}\zeta^{r-1-j}+U_{r-N}(\zeta). \tag{2.20}$$

Now $U_{r-N}(\zeta)$ satisfies (2.14) with r replaced by $r-N$. By (2.16), we conclude that

$$\begin{aligned}&\int_0^\infty e^{-rt}\varphi(t)\,dt\sim\varphi_0(\sigma-1)!/r^\sigma && \text{as } r\to\infty,\\ &\int_0^\infty e^{-rt}|\varphi(t)|\,dt\sim|\varphi_0|(\sigma-1)!/r^\sigma && \text{as } r\to\infty,\end{aligned} \tag{2.21}$$

see [6, p. 81]. Also, $g_{p,i}(r)=O(r^p)$ as $r\to\infty$, by Lemma 2.1. Consequently, for $\zeta\in S(\rho,\theta_0)$

$$|U_{r-N}(\zeta)|\leqq 0\left(r^{p-\sigma}|\zeta|^{r-N}\prod_{j=1}^{m}\left[\mu_j(r-N+p)+\nu_j\right]!\right) \quad \text{as } r\to\infty, \tag{2.22}$$

uniformly in ζ. Similarly

$$a_{r-N+p}\sim\varphi_0 r^{p-\sigma}\prod_{j=1}^{m}\left[\mu_j(r-N+p)+\nu_j\right]! \quad \text{as } r\to\infty. \tag{2.23}$$

Comparing (2.22) and (2.23), we obtain

$$|U_{r-N}(\zeta)|\leqq 0\left(a_{r-N+p}|\zeta|^{r-N}\right) \quad \text{as } r\to\infty, \tag{2.24}$$

uniformly in ζ for all $\zeta\in S(\rho,\theta_0)$.

Finally, by choosing $N=k+p+1$, and recalling that (see (2.23) above)

$$\lim_{r\to\infty}\frac{a_{r-j}}{a_r}=0, \qquad j=1,2,\cdots, \tag{2.25}$$

(2.17) follows. □

COROLLARY 2.4. *If* $\sum_{i=1}^{m}\mu_i=\mu$, *where* μ *is a positive integer, and if* $\varphi(t)$ *also satisfies*

$$\varphi(t)\sim\sum_{i=0}^{\infty}\varphi_i t^{i+\sigma-1}\quad \text{as } t\to 0^+,\quad \text{for some } \sigma>0, \tag{2.26}$$

then $U_r(\zeta)$ *is of the form*

$$U_r(\zeta)\sim a_r\zeta^r\sum_{i=0}^{\infty}\frac{\beta_i(\zeta)}{r^{i+\mu}}\quad \text{as } r\to\infty, \tag{2.27}$$

with $\beta_i(\zeta)$ *being polynomials in* ζ^{-1}. *Furthermore* (2.27) *is uniformly valid in* ζ *for* $\zeta\in T=S(\rho,\theta_0)\backslash\{\zeta\colon |\zeta|<\varepsilon\}$, *for any* $\varepsilon>0$.

Proof. As mentioned in the remark following the statement of Theorem 2.1, (2.4) and (2.26) imply that $w(r)\sim\sum_{i=0}^{\infty}\varphi_i(i+\sigma-1)!/r^{i+\sigma}$ as $r\to\infty$. This, together with the result

$$x^{\beta-\alpha}\frac{(x+\alpha)!}{(x+\beta)!}\sim 1+\sum_{i=1}^{\infty}\frac{c_i}{x^i}\quad \text{as } x\to\infty, \tag{2.28}$$

c_i being some constants independent of x (see [1, p. 257, formula 6.1.47]), give

$$\frac{a_{r-1-j}}{a_r}\sim r^{-\mu(1+j)}\left[1+\sum_{i=1}^{\infty}\frac{d_i^{(j)}}{r^i}\right]\quad \text{as } r\to\infty, \tag{2.29}$$

where $d_i^{(j)}$ are constants independent of r. Upon substituting (2.29) in (2.17), we obtain

$$U_r(\zeta)=a_r\zeta^r\left[\sum_{i=0}^{\mu(k+1)-1}\frac{\beta_i(\zeta)}{r^{i+\mu}}+O\left(\frac{1}{r^{\mu(k+1)}}\right)\right]\quad \text{as } r\to\infty, \tag{2.30}$$

with $\beta_i(\zeta)$ being given by

$$\beta_{j\mu+i}(\zeta)=-\sum_{l=0}^{j}d_{l\mu+i}^{(j-l)}\zeta^{-j+l-1},\quad 0\leqq i\leqq\mu-1,\quad j=0,1,\cdots, \tag{2.31}$$

where $d_0^{(j)}=1$, $j=0,1,\cdots$; hence $\beta_0(\zeta)=-\zeta^{-1}$. This completes the proof of the corollary. □

Remark. Under the conditions stated in the corollary above, we have actually shown that the partial sums of the everywhere divergent series $F(\zeta):=\sum_{r=1}^{\infty}a_r\zeta^r$ are of the form given in (1.7) and (1.8), with $A(\zeta)=\mathscr{F}(\zeta)$, the Borel-type sum of $F(\zeta)$, and $g_i(\zeta)=0$, $0\leqq i\leqq\mu-1$.

As an example, consider one of the series given in [11, Table A3], namely $\sum_{r=1}^{\infty}(-1)^{r-1}c_r/x^r$, with $c_1=2$ and $c_r=c_{r-1}(2r-3)^2$, $r\geqq 2$. Therefore, $(-1)^{r-1}c_r/x^r=a_r\zeta^r$, with $\zeta=-4/x$ and $a_r=-[(r-3/2)!]^2/(2\pi)$, $r\geqq 1$. That is to say, $\mu_1=\mu_2=1$, $\nu_1=\nu_2=-3/2$, and $r^p w(r)=Cr^{p'}$ with $C=-1/(2\pi)$ and $p'=p-1=0$ in Remark (4) following (2.6b). Consequently,

$$\mathscr{F}(\zeta)=-\frac{\zeta}{2\pi}\int_0^{\infty}\int_0^{\infty}e^{-(t_1+t_2)}\frac{(t_1t_2)^{-1/2}}{1-\zeta t_1t_2}\,dt_1\,dt_2.$$

By making the transformation of variables $t_1=\sqrt{x}\,t\cos^2\theta$, $t_2=\sqrt{x}\,t\sin^2\theta$, and performing the integral with respect to θ, we obtain

$$\mathscr{F}(\zeta)=\frac{2}{\sqrt{x}}\int_0^\infty e^{-\sqrt{x}t}\frac{dt}{(1+t^2)^{1/2}}=\frac{\pi}{\sqrt{x}}\left[\mathbf{H}_0(\sqrt{x})-Y_0(\sqrt{x})\right],$$

where $\mathbf{H}_\nu(z)$ is the Struve function of order ν [1, p. 496, formula 12.1.8]. The numerical result given in [11] for $x=4$ is another indication that the u-transformation produces approximations to $\mathscr{F}(\zeta)$.

Finally, we note that when $\mu_i=\nu_i=0$, $1\leqq i\leqq m$, in Theorem 2.2, $\mathscr{F}(\zeta)$ converges for $|\zeta|<1$ and diverges for $|\zeta|>1$. Thus, $\mathscr{F}(\zeta)$ represents an analytic function $u(\zeta)$ within the unit circle. Equation (2.6) now becomes

$$\mathscr{F}(\zeta)=\int_0^\infty \varphi(t)\left[z\frac{d}{dz}\right]^p\frac{z}{1-z}\,dt,\quad z=\zeta e^{-t}. \tag{2.32}$$

Since this time $\mathscr{F}(\zeta)$ is analytic in the ζ-plane cut along $[1,\infty)$, it represents the analytic continuation of $u(\zeta)$ outside the unit circle. Furthermore, (2.10) holds with (2.11) replaced by

$$U_r(\zeta)=\int_0^\infty \varphi(t)Q_p(r,z)\,dt. \tag{2.33}$$

Let us now assume that $\varphi(t)$ satisfies (2.26). Then substituting (2.12) in (2.33), and applying Watson's lemma for $r\to\infty$, after some manipulation of the asymptotic expansions that arise, we obtain (2.27) with $\mu=0$ there. Of course, in this case the $\beta_i(\zeta)$ are not necessarily polynomials in ζ^{-1}. In addition, (2.27) with $\mu=0$ holds for all $\zeta\notin[1,\infty)$, for which $F(\zeta)$ converges or diverges. The details are left to the interested reader.

3. Further developments. The results of the previous section have been based mainly on the assumptions of Theorem 2.2, namely (2.4) to (2.4b). It is these assumptions that enable one to express the Borel-type sum $\mathscr{F}(\zeta)$ of $F(\zeta)$ as in (2.6) to (2.6b). One important feature of (2.6) is the function $Q_p(1,z)=(zd/dz)^p(z/(1-z))$, which is very easy to handle. Actually this function has simple expansions about $z=0$ and $z=\infty$, and it is these expansions that lead to the results of Theorem 2.2, Theorem 2.3, and Corollary 2.4. In this section we seek to generalize the conditions of Theorem 2.2 in a way that will enable us to retain the function $Q_p(1,z)$. We note that the developments of this section can readily be applied to generalized hypergeometric functions.

THEOREM 3.1. *Let a_r be expressible in the form*

$$a_r=r^P w(r)\prod_{i=1}^{m}(\mu_i r+\nu_i)!\prod_{j=1}^{n}B\left(\kappa_j r+\lambda_j+1,\bar{\kappa}_j r+\bar{\lambda}_j+1\right), \tag{3.1}$$

where p, $w(r)$, μ_i, and ν_i are exactly as in Theorem 2.2,

$$\kappa_j\geqq 0,\quad \bar{\kappa}_j\geqq 0,\quad \kappa_j+\bar{\kappa}_j>0,\quad \kappa_j+\lambda_j>-1,\quad \bar{\kappa}_j+\bar{\lambda}_j>-1,\qquad j=1,\cdots,n, \tag{3.1a}$$

and $B(b,c)$ is the beta function defined by

$$B(b,c)=\int_0^1 \tau^{b-1}(1-\tau)^{c-1}d\tau=\frac{(b-1)!(c-1)!}{(b+c-1)!},\qquad \mathrm{Re}\,b>0,\qquad \mathrm{Re}\,c>0. \tag{3.1b}$$

It is clear that the power series $F(\zeta) := \sum_{r=1}^{\infty} a_r \zeta^r$ *diverges for all* $\zeta \neq 0$. *Define* $S(\rho, \theta_0)$ *again as in Theorem* 2.2. *Then* $F(\zeta)$ *is the asymptotic expansion of its Borel-type sum*

$$\mathscr{F}(\zeta) = \underbrace{\int_0^{\infty} \cdots \int_0^{\infty}}_{m+1 \text{ times}} \underbrace{\int_0^{1} \cdots \int_0^{1}}_{n \text{ times}} \psi(\vec{t}, \vec{\tau}) Q_p(1, z) \prod_{i=0}^{m} dt_i \prod_{j=1}^{n} d\tau_j, \tag{3.2}$$

as $\zeta \to 0$, *uniformly in* $S(\rho, \theta_0)$, *for each finite* ρ, *where* $\vec{t} = (t_0, t_1, \cdots, t_m)$, $\vec{\tau} = (\tau_1, \cdots, \tau_n)$, *and*

$$\psi(\vec{t}, \vec{\tau}) = \exp\left(- \sum_{i=1}^{m} t_i \right) \left(\prod_{i=1}^{m} t_i^{\nu_i} \right) \varphi(t_0) \left(\prod_{j=1}^{n} \left[\tau_j^{\lambda_j} (1 - \tau_j)^{\bar{\lambda}_j} \right] \right), \tag{3.2a}$$

$$z = \zeta e^{-t_0} \left(\prod_{i=1}^{m} t_i^{\mu_i} \right) \left(\prod_{j=1}^{n} \left[\tau_j^{\kappa_j} (1 - \tau_j)^{\bar{\kappa}_j} \right] \right), \tag{3.2b}$$

with the integrals over t_i, $0 \leqq i \leqq m$ *being from* 0 *to* ∞ *and those over* τ_j, $1 \leqq j \leqq n$, *from* 0 *to* 1. *The function* $\mathscr{F}(\zeta)$ *is analytic in the* ζ*-plane cut along* $[0, \infty)$.

Proof. Similar to that of Theorem 2.2. □

Remark. If in (3.1) $r^p w(r) = C r^{p'}$, where C is a constant and $p' = p - 1 \geqq 0$, then the Borel-type sum of $\mathscr{F}(\zeta)$ in (3.2) reduces to that obtained from (3.2) by omitting the integration with respect to t_0 after $\varphi(t_0)$ in $\psi(\vec{t}, \vec{\tau})$ has been replaced by C, and p by p', and the factor e^{-t_0} has been deleted from z. (cf. Remark (4) following statement of Theorem 2.2.)

THEOREM 3.2. *Assume that all the conditions of Theorem* 3.1 *are satisfied, and, in addition,* $\varphi(t)$ *is as in Theorem* 2.3. *Then, for any integer* $k \geqq 0$,

$$U_r(\zeta) = \mathscr{F}(\zeta) - A_{r-1}(\zeta) = - \sum_{j=0}^{k-1} a_{r-1-j} \zeta^{r-1-j} + O\left(a_{r-1-k} \zeta^{r-1-k-p} \right) \quad \text{as } r \to \infty, \tag{3.3}$$

uniformly in ζ *for* $\zeta \in S(\rho, \theta_0)$ *for each finite* ρ.

Proof. Similar to that of Theorem 2.3. □

COROLLARY 3.3. *If* $\sum_{i=1}^{m} \mu_i = \mu$, *where* μ *is an integer, and if* $\varphi(t)$ *is as in Corollary* 2.4, *then* $U_r(\zeta)$ *is of the form given in* (2.27), *with* $\beta_i(\zeta)$ *being polynomials in* ζ^{-1} *again.* (2.27) *is uniformly valid in* ζ *for* $\zeta \in T$.

Proof. Using Stirling's formula, it can be shown that for $r \to \infty$

$$B(\kappa r + \lambda, \bar{\kappa} r + \bar{\lambda}) \sim \begin{cases} \left[\dfrac{\kappa^{\kappa} \bar{\kappa}^{\bar{\kappa}}}{(\kappa + \bar{\kappa})^{\kappa + \bar{\kappa}}} \right]^r \displaystyle\sum_{i=0}^{\infty} \frac{e_i}{r^{i+1/2}} & \text{if } \kappa > 0, \quad \bar{\kappa} > 0, \\ \displaystyle\sum_{i=0}^{\infty} \frac{e_i'}{r^{i+\bar{\lambda}}} & \text{if } \kappa > 0, \quad \bar{\kappa} = 0, \end{cases} \tag{3.4}$$

where e_i and e_i' are constants independent of r. With the help of (3.4), the proof of this corollary can now be accomplished as that of Corollary 2.4. □

Note. The results above are applicable to series $F(\zeta)$ for which

$$a_r = r^p w(r) \frac{\prod_{i=1}^{m'} (\varepsilon_i r + \delta_i)!}{\prod_{i=1}^{n'} (\varepsilon_i r + \bar{\delta}_i)!}, \tag{3.5}$$

with p and $w(r)$ as before, $\varepsilon_i > 0$, $\varepsilon_i + \delta_i > -1$, $\bar{\delta}_i > \delta_i$, $1 \leqq i \leqq n'$, $m' > n'$. This is so since a_r can be expressed as in (3.1), due to the fact that

$$\frac{(\varepsilon r+\delta)!}{(\varepsilon r+\bar{\delta})!} = \frac{B(\varepsilon r+\delta+1, \bar{\delta}-\delta)}{(\bar{\delta}-\delta-1)!}. \tag{3.6}$$

There is no loss of generality in assuming that $\bar{\delta}_i > \delta_i$, $1 \leqq i \leqq n'$, for if $\bar{\delta}_i \leqq \delta_i$ for some i, say $i = q$, then $(\varepsilon_q r + \bar{\delta}_q)!$ and r^p in (3.5) can be replaced by $[\varepsilon_q r + (k\varepsilon_q + \bar{\delta}_q)]!$ and the polynomial $r^p \prod_{j=1}^{k} (\varepsilon_q r + j\varepsilon_q + \bar{\delta}_q)$ respectively, such that $\hat{\delta}_q = k\varepsilon_q + \bar{\delta}_q > \delta_q$. In general, we can express a_r as $a_r = \sum_{j=1}^{k} h_j a_r^{(j)}$, where

$$a_r^{(j)} = r^{p+j} w(r) \frac{\prod_{i=1}^{m'} (\varepsilon_i r + \delta_i)!}{\prod_{i=1}^{n'} (\varepsilon_i r + \hat{\delta}_i)!}, \qquad 1 \leqq j \leqq k,$$

with $\hat{\delta}_i > \delta_i$, $1 \leqq i \leqq n'$. Now we apply Theorem 3.1, Theorem 3.2 and Corollary 3.3 to each of the series $\sum_{r=1}^{\infty} h_j a_r^{(j)} \zeta^r$ and add the results. The overall result is that Theorem 3.1, Theorem 3.2 and Corollary 3.3 hold for the series $F(\zeta) := \sum_{r=1}^{\infty} a_r \zeta^r$, even though $\bar{\delta}_i > \delta_i$ is not satisfied for all $1 \leqq i \leqq n'$. Thus our results can be applied to the generalized hypergeometric functions ${}_pF_q$, where

$${}_pF_q\left(\begin{matrix} \alpha_1, \cdots, \alpha_p \\ \rho_1, \cdots, \rho_q \end{matrix} \middle| \zeta \right) = \sum_{k=0}^{\infty} \frac{\prod_{h=1}^{p} (\alpha_h)_k \zeta^k}{\prod_{h=1}^{q} (\rho_h)_k k!}$$

with $(c)_k = \Gamma(c+k)/\Gamma(c)$, $k = 0, 1, \cdots$, see [5, p. 155], when $p > q+1$.

Note also that the representation given in (3.2) is somewhat related to the beta transform described in [5, p. 160].

As an example, we consider the asymptotic series $\sum a_r \zeta^r$ with

$$a_r = (\alpha)_{r-1} (1+\alpha-\beta)_{r-1} / (n-1)! \quad \text{and} \quad \zeta = -1/x.$$

That is to say, a_r is expressible as

$$a_r = -\frac{(r+\alpha-\beta-1)! B(r+\alpha-1, 1-\alpha)}{(-\alpha)!(\alpha-1)!(\alpha-\beta)!}.$$

By the remark above, $\mathscr{F}(\zeta)$ can be expressed as a double integral that can be reduced to a one-dimensional integral, which, by using some relations among the confluent hypergeometric functions of different parameters, can be shown to be $-x^{\alpha-1} U(\alpha, \beta, x)$. Again [11, Table A3] contains numerical results for different values of α, β, and x, that indicate that the u-transformation produces approximations to $\mathscr{F}(\zeta)$.

4. Concluding remarks. We have shown that under the conditions stated in Corollary 2.4 and Corollary 3.3, the partial sums $A_{r-1}(\zeta) = \sum_{i=1}^{r-1} a_i \zeta^i$ of the everywhere divergent series $F(\zeta) := \sum_{i=1}^{\infty} a_i \zeta^i$ are of the form (1.7) and (1.8), where $A(\zeta)$ is the Borel-type sum of $F(\zeta)$. As mentioned in the introduction to this work, most of the examples of everywhere divergent series considered in [11] satisfy the requirements of Corollary 2.4 and Corollary 3.3; furthermore, after some tedious calculations, involving manipulation of (2.6) and (3.2), one observes for all these examples, that the numbers obtained by applying Levin's t or u transformation to $A_i(\zeta)$, are approximations to the Borel-type sum of $F(\zeta)$. In view of this observation, we conjecture that for the kind of series considered in Corollary 2.4 and Corollary 3.3, Levin's t and u transformations produce approximations that converge to the Borel-type sums of these series.

Acknowledgment. This work was done at the NASA Lewis Research Center while the author was a National Research Council associate.

REFERENCES

[1] M. ABRAMOWITZ AND I. A. STEGUN, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, Nat. Bur. Standards Appl. Math. Series No. 55, Superintendent of Documents, U. S. Government Printing Office, Washington, DC, 1964.

[2] R. B. DINGLE, *Asymptotic Expansions: Their Derivation and Interpretation*, Academic Press, New York, 1973.

[3] W. B. FORD, *Studies on Divergent Series and Summability and the Asymptotic Developments of Functions Defined by Maclaurin Series*, Chelsea, Bronx, New York, 1960.

[4] D. LEVIN, *Development of non-linear transformations for improving convergence of sequences*, Internat. J. Comput. Math., B3 (1973), pp. 371–388.

[5] Y. L. LUKE, *Mathematical Functions and their Approximations*, Academic Press, New York, 1975.

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

[7] A. SIDI, *Convergence properties of some non-linear sequence transformations*, Math. Comp., 33 (1979), pp. 315–326.

[8] _____, *Analysis of convergence of the T-transformation for power series*, Math. Comp., 35 (1980), pp. 833–850.

[9] _____, *Converging factors for some asymptotic moment series that arise in numerical quadrature*, J. Austral. Math. Soc. (Series B), 24 (1982), pp. 223–233.

[10] D. A. SMITH AND W. F. FORD, *Acceleration of linear and logarithmic convergence*, SIAM J. Numer. Anal., 16 (1979), pp. 223–240.

[11] D. A. SMITH AND W. F. FORD, *Numerical comparisons of nonlinear convergence accelerators*, Math. Comp., 38 (1982), pp. 481–499.

SIAM J. MATH. ANAL.
Vol. 17, No. 5, September 1986

013

NIELSEN'S GENERALIZED POLYLOGARITHMS*

K. S. KÖLBIG†

Abstract. Properties (in particular functional relations and special values) of the functions

$$(-1)^{n+p-1}(n-1)!p!S_{n,p}(z)=\int_0^1 \log^{n-1}t\log^p(1-zt)\frac{dt}{t},$$

$$(-1)^{n+p-1}(n-1)!p!L_{n,p}(z)=\int_0^z \log^{n-1}t\log^p(1-t)\frac{dt}{t},$$

$$(-1)^{n-1}(n-1)!p!M_{n,p}(z)=\int_0^z \log^{n-1}t\log^p(1+t)\frac{dt}{t},$$

which play a role in the computation of higher order radiative corrections in quantum electrodynamics, are discussed for complex z and positive integers n and p. The first function is a generalization of the well-known polylogarithms ($p=1$). The discussion is based on results published by Nielsen early this century in a little-known monograph.

Key words. Nielsen's generalized polylogarithms, polylogarithms, Spence functions, logarithmic integrals, Riemann zeta functions, Stirling numbers of the first kind

AMS(MOS) subject classification. Primary 33A70

1. Introduction.

A certain class of logarithmic integrals, the so-called polylogarithms, defined by

$$Li_n(x)=\frac{(-1)^{n-1}}{(n-2)!}\int_0^1 \log^{n-2}t\log(1-xt)\frac{dt}{t}, \qquad (n\geqq 2), \tag{1.1}$$

has been investigated in the past by many mathematicians, including Euler, Kummer, Abel and Spence. In particular, their properties with respect to transformations of the variable were established for several special cases, for example, for the dilogarithm ($n=2$) and the trilogarithm ($n=3$). A special property of these functions is the relation

$$Li_n(1)=\zeta(n) \tag{1.2}$$

where $\zeta(n)$ is the Riemann zeta function of integer argument.

At the beginning of this century, the Danish mathematician Niels Nielsen collected the known results into a monograph [20], and introduced a new class of functions, namely

$$S_{n,p}(z)=\frac{(-1)^{n+p-1}}{(n-1)!p!}\int_0^1 \log^{n-1}t\log^p(1-zt)\frac{dt}{t}, \tag{1.3}$$

defined for positive integers n,p, and complex $z=x+iy$. The function $\log\zeta$ is understood to lie on its principal sheet, and we define

$$\log\xi=\lim_{\varepsilon\to 0+}\log(\xi+i\varepsilon)=\log|\xi|+i\pi \quad \text{if } \xi<0. \tag{1.4}$$

* Received by the editors September 10, 1984.
† European Organization for Nuclear Research (CERN), CH-1211 Geneva 23, Switzerland.

Because of the fact that

$$S_{n-1,1}(z)=Li_n(z),$$

these functions are generalizations of the polylogarithms (1.1). It is likely that Nielsen undertook his research on $S_{n,p}(z)$ in the hope of finding expressions for $\zeta(2k+1)$ similar to those known for $\zeta(2k)$, and results for

$$\beta(2k)=\sum_{j=0}^{\infty}\frac{(-1)^j}{(2j+1)^{2k}} \qquad (k=1,2,\cdots), \tag{1.5}$$

similar to those known for $\beta(2k+1)$. At least he remarked [20, pp. 131, 204] that he was not able to obtain such results from his relations for $S_{n,p}(z)$.

Nielsen's monograph [20] contains many formulae for $S_{n,p}(z)$ and for two related functions $L_{n,p}(z)$ and $M_{n,p}(z)$, in particular involving transformations of the argument z. Although most of his formulae are correct (apart from misprints) for complex $z=x+iy$ with $y\neq 0$, the direct replacement of z by real x may lead to discrepancies when the functions involved are complex-valued for real x. This fact has been overlooked in [12]. As far as the author knows, these formulae have never found their way into any of the relevant handbooks, and it seems that the only reference to this monograph until about fifteen years ago was in the book of Lewin [16]. About this time, however, it became apparent that Nielsen, although he did not succeed in finding results for $\zeta(2k+1)$ and $\beta(2k)$, had introduced a class of functions which are of importance in certain applications, in particular in quantum electrodynamics (see, for example, [2], [3], [6], [15], [18], [19]). With this application in mind, functional relations for $S_{n,p}(z)$ were presented in [12], together with a method for computing $S_{n,p}(z)$ accurately (12–14 digits) for real $z=x$ and $n+p\leqq 5$. Also some of Nielsen's ideas, in particular on the special values $s_{n,p}=S_{n,p}(1)$ and on a similar integral, have been investigated further [12], [13], [14]. Jacobs and Lambert [11] have developed a method for computing $S_{1,1}(z)$ and $S_{2,1}(z)$ for complex values of z, and Barlow [4] has considered the computation of $S_{n,p}(z)$ for complex z by continued fraction approximants.

An explanation for the fact that Nielsen's monograph remained undiscovered for so long lies perhaps in the fact that it was published in a journal which (at least for mathematicians and physicists) was little-known and not easy to obtain. Further, it contains an unusually large number of misprints and, as we shall see, genuine errors. It is, however, far from being out-of-date or without interest. Integrals of this type or even, in some respects, more general, such as

$$\int_0^1 \frac{\log^m t\log^n(1-t)\log^p(1+t)}{D(t)}\,dt$$

where m,n, and p are nonnegative integers, and $D(t)=t$, $1-t$, or $1+t$, play a role in the evaluation of Feynman and relativistic phase space integrals, as is indicated by the recent publication of a table by Gastmans and Troost [7]. These authors also discuss certain special values of the polylogarithms, a problem which Nielsen investigated in a more general way for his functions, and give a table of infinite series related to polylogarithmic integrals.

Therefore, for historical and practical reasons, it is perhaps worthwhile to give a systematic and critical presentation of the more important part of Nielsen's work on generalized polylogarithms, which until now has not been generally accessible. For

example, Lewin [17, p. 199], in a new edition of [16], remarked that it was not possible for him to give a discussion of the general results obtained by Nielsen, although he cites Nielsen's monograph many times. Also, the table of Gröbner and Hofreiter [9, pp. 71–73] contains a chapter on "Euler's dilogarithm and its generalisations" and a reference to [20], but only formulae for $p=1$ are given there.

In the presentation which follows, we shall follow the original monograph fairly closely, but on some occasions we shall present new results or proofs; for example, certain constants appearing in the functional equation relating $S_{n,p}(z)$ to $S_{\nu,\rho}(1/z)$ will be expressed in terms of known constants. Special attention will also be given to the correct representation of the functional relations for real arguments.

2. Basic formulae. Nielsen's generalized polylogarithms can be defined by [12], [20]

$$S_{n,p}(z)=\frac{(-1)^{n+p-1}}{(n-1)!p!}\int_0^1 \log^{n-1}t\log^p(1-zt)\frac{dt}{t} \tag{2.1}$$

$$=\sum_{j=0}^{\infty}\frac{(-1)^j S_{p+j}^{(p)}}{(p+j)!(p+j)^n}z^{p+j} \qquad (|z|\leqq 1) \tag{2.2}$$

$$=\frac{(-1)^{n-1}}{(n-1)!p!}\left[\frac{\partial^{n+p-1}}{\partial\beta^{n-1}\partial\alpha^p}\frac{1}{\beta}{}_2F_1(\alpha,\beta;\beta+1;z)\right]_{\alpha=\beta=0} \tag{2.3}$$

where

$${}_2F_1(\alpha,\beta;\gamma;z)=\frac{\Gamma(\gamma)}{\Gamma(\beta)\Gamma(\gamma-\beta)}\int_0^1 t^{\beta-1}(1-t)^{\gamma-\beta-1}(1-tz)^{-\alpha}dt$$

is the Gaussian hypergeometric function, and

$$S_k^{(m)}=\sum_{i=0}^{k-m}\frac{1}{i!}\binom{k-1+i}{k-m+i}\binom{2k-m}{k-m-i}\sum_{j=0}^{i}(-1)^i\binom{i}{j}^{k-m+i}$$

are the Stirling numbers of the first kind in Schlömilch's representation [5, p. 216], generated by [1], [5, p. 212]

$$\log^m(1+x)=m!\sum_{k=m}^{\infty}S_k^{(m)}\frac{x^k}{k!} \qquad (|x|<1). \tag{2.4}$$

In this paper, we shall use Nielsen's notation

$$S_n(z)=S_{n-1,1}(z)$$

instead of the notation $Li_n(z)$ introduced by Lewin [16], [17].

It follows from the series definition (2.2) that, for the polylogarithms,

$$S_n(z)=\sum_{j=1}^{\infty}\frac{z^j}{j^n}=\frac{z}{1^n}+\frac{z^2}{2^n}+\frac{z^3}{3^n}+\cdots \qquad (|z|\leqq 1)\quad (n>1). \tag{2.5}$$

For $n=1$ and $n=0$, we have as degenerate cases

$$S_1(z)=-\log(1-z),\quad S_0(z)=\frac{z}{1-z},\quad (|z|<1). \tag{2.6}$$

For $p>1$, the series (2.2) becomes considerably more complicated. By writing [5, p. 217]

$$S_{j+2}^{(2)}=(-1)^j(j+1)!\sum_{k=1}^{j+1}\frac{1}{k}, \tag{2.7}$$

$$S_{j+3}^{(3)}=(-1)^j(j+2)!\left[\left(\sum_{k=1}^{j+2}\frac{1}{k}\right)^2-\sum_{k=1}^{j+2}\frac{1}{k^2}\right], \tag{2.8}$$

we find, for example

$$S_{n,2}(z)=\frac{z^2}{2^{n+1}}+\left(1+\frac{1}{2}\right)\frac{z^3}{3^{n+1}}+\left(1+\frac{1}{2}+\frac{1}{3}\right)\frac{z^4}{4^{n+1}}+\cdots, \tag{2.9}$$

$$S_{n,3}(z)=\frac{1}{2}\left[\left(1+\frac{1}{2}\right)^2-\left(1+\frac{1}{4}\right)\right]\frac{z^3}{3^{n+1}} + \frac{1}{2}\left[\left(1+\frac{1}{2}+\frac{1}{3}\right)^2-\left(1+\frac{1}{4}+\frac{1}{9}\right)\right]\frac{z^4}{4^{n+1}}+\cdots. \tag{2.10}$$

From (2.1) and (2.2) one may easily obtain

$$\frac{d}{dz}S_{n,p}(\alpha z)=\begin{cases}\dfrac{(-1)^p}{p!}\dfrac{1}{z}\log^p(1-\alpha z), & (n=1),\\[2ex] \dfrac{1}{z}S_{n-1,p}(\alpha z), & (n\geqq 2),\end{cases} \tag{2.11}$$

$$\begin{aligned} S_{n,p}(\alpha z)&=\int_0^z\frac{1}{\zeta}S_{n-1,p}(\alpha\zeta)\,d\zeta, \qquad (n\geqq 1),\\ S_{n,p}(z)&=\frac{1}{p!p^n}z^p+O(z^{p+1}), \qquad (z\to 0),\end{aligned} \tag{2.12}$$

and from (5.12) below,

$$S_{n,p}(z)=\frac{(-1)^p}{(n+p)!}\log^{n+p}(-z)+O\left(\log^{n-1}(-z)\right), \quad (z\to\infty).$$

In order to derive his relations, Nielsen introduces in addition the functions

$$L_{n,p}(z)=\frac{(-1)^{n+p-1}}{(n-1)!p!}\int_0^z\log^{n-1}t\log^p(1-t)\frac{dt}{t}, \qquad (n\geqq 1, p\geqq 1), \tag{2.13}$$

$$M_{n,p}(z)=\frac{(-1)^{n-1}}{(n-1)!p!}\int_0^z\log^{n-1}t\log^p(1+t)\frac{dt}{t}, \qquad (n\geqq 1, p\geqq 1). \tag{2.14}$$

Using the binomial theorem, it is easy to see that

$$S_{n,p}(z)=\sum_{j=0}^{n-1}\frac{\log^j z}{j!}L_{n-j,p}(z), \tag{2.15}$$

$$L_{n,p}(z)=\sum_{j=0}^{n-1}\frac{(-1)^j\log^j z}{j!}S_{n-j,p}(z), \tag{2.16}$$

and

$$(-1)^p S_{n,p}(-z)=\sum_{j=0}^{n-1}\frac{\log^j z}{j!}M_{n-j,p}(z), \tag{2.17}$$

$$M_{n,p}(z)=(-1)^p\sum_{j=0}^{n-1}\frac{(-1)^j\log^j z}{j!}S_{n-j,p}(-z). \tag{2.18}$$

In particular,

$$L_{1,p}(z)=S_{1,p}(z),\qquad M_{1,p}(z)=(-1)^p S_{1,p}(-z). \tag{2.19}$$

The values of $S_{n,p}(z)$ for $z=1,-1$ and $\frac{1}{2}$ are of special interest. These values are real, and we define

$$\begin{aligned} s_{n,p}&=S_{n,p}(1)=L_{n,p}(1),\\ \sigma_{n,p}&=(-1)^p S_{n,p}(-1)=M_{n,p}(1),\\ a_{n,p}&=S_{n,p}(\tfrac{1}{2}). \end{aligned} \tag{2.20}$$

We also write

$$s_n=s_{n-1,1},\quad \sigma_n=\sigma_{n-1,1},\quad a_n=a_{n-1,1}.$$

Properties of $s_{n,p}$, $\sigma_{n,p}$ and $a_{n,p}$ are discussed in §9. For the moment, we note only that $s_{n,p}$ can be expressed in $\zeta(k)$, that

$$s_{n,p}=s_{p,n}, \tag{2.21}$$

which follows from the definition integral (2.1) by partial integration, and that

$$s_n=\zeta(n),\qquad \sigma_n=(1-2^{1-n})\zeta(n), \tag{2.22}$$

which follows from (2.5), using well-known properties of the Riemann zeta function.

3. Some analytical properties of $S_{n,p}(z)$, $L_{n,p}(z)$, and $M_{n,p}(z)$. With the general definition of the logarithm function, $S_{n,p}(z)$, $L_{n,p}(z)$, and $M_{n,p}(z)$ are multi-valued functions of z. By restricting $\log\zeta$ to its principal sheet, using (1.4), and cutting the z-plane appropriately, we obtain single-valued functions which are holomorphic at all finite points z with the exception of the branch-cuts. In the remainder of this section, we define these cuts along sections of the real axis $z=x$ and describe the behavior of these functions on this axis, in particular on the cuts. These properties will be required later.

3.1. $S_{n,p}(z)$. In this case, we have as branch points $z=1$ and $z=\infty$, and we cut the z-plane along the real axis from $x=1$ to ∞. For $x\leqq 1$, $S_{n,p}(x)$ is real. For the behavior on the cut, we have from (1.3) and (1.4)

$$\begin{aligned} &\lim_{\varepsilon\to 0+} S_{n,p}(x-i\varepsilon)=S_{n,p}(x),\\ &\lim_{\varepsilon\to 0+} S_{n,p}(x+i\varepsilon)=S^*_{n,p}(x), \end{aligned}\qquad (x>1), \tag{3.1}$$

where the asterisk denotes the complex conjugate. Therefore, $S_{n,p}(z)$ is, on the cut, continuous from below.

3.2. $L_{n,p}(z)$ $(n>1)$. Because of (2.19), we can make the restriction $n>1$. As branch points, we find $z=0$, $z=1$ and $z=\infty$, and we cut the z-plane along the real axis from $x=1$ to ∞, and from $x=0$ to $-\infty$. For $0\leqq x\leqq 1$, $L_{n,p}(x)$ is real. On the cuts, we have from (2.16),

$$\begin{aligned} &\lim_{\varepsilon\to 0+} L_{n,p}(x+i\varepsilon)=L_{n,p}(x),\\ &\lim_{\varepsilon\to 0+} L_{n,p}(x-i\varepsilon)=L^{*}_{n,p}(x), \end{aligned} \qquad (x<0), \tag{3.2}$$

and

$$\begin{aligned} &\lim_{\varepsilon\to 0+} L_{n,p}(x-i\varepsilon)=L_{n,p}(x),\\ &\lim_{\varepsilon\to 0+} L_{n,p}(x+i\varepsilon)=L^{*}_{n,p}(x), \end{aligned} \qquad (x>1), \tag{3.3}$$

so that $L_{n,p}(z)$ is continuous from above for $x<0$, and continuous from below for $x>1$.

3.3. $M_{n,p}(z)$. From (2.19) and (3.1), we find for $n=1$ that the branch points are $z=-1$ and $z=\infty$, and we cut the z-plane from $x=-1$ to $-\infty$. On the cut, we have

$$\begin{aligned} &\lim_{\varepsilon\to 0+} M_{1,p}(x+i\varepsilon)=M_{1,p}(x),\\ &\lim_{\varepsilon\to 0+} M_{1,p}(x-i\varepsilon)=M^{*}_{1,p}(x), \end{aligned} \qquad (x<-1), \tag{3.4}$$

so that $M_{1,p}(z)$ is continuous from above for $x<-1$. For $x\geqq -1$, $M_{1,p}(x)$ is real.

For $n>1$, we find from (2.18)

$$\lim_{\varepsilon\to 0+} M_{n,p}(x\pm i\varepsilon)=(-1)^{p}\sum_{j=0}^{n-1}\frac{(-1)^{j}\log^{j}(x\pm i\varepsilon)}{j!}S_{n-j,p}(-x\mp i\varepsilon).$$

Because of (1.4) and (3.1), it follows that $M_{n,p}(z)$ has branch points at $z=0$ and $z=\infty$, and we cut the z-plane from $x=0$ to $-\infty$. Then

$$\begin{aligned} &\lim_{\varepsilon\to 0+} M_{n,p}(x+i\varepsilon)=M_{n,p}(x),\\ &\lim_{\varepsilon\to 0+} M_{n,p}(x-i\varepsilon)=M^{*}_{n,p}(x), \end{aligned} \qquad (x<0). \tag{3.5}$$

For $x\geqq 0$, $M_{n,p}(x)$ is real. Thus, $M_{n,p}(z)$ is continuous from above for $x<0$ $(n\geqq 1, p\geqq 1)$.

4. A lemma. In order to establish some of the functional relations for the generalized polylogarithms, Nielsen needs the following

LEMMA. *Let the complex quantities* x_j, y_j $(j=1,\cdots,n)$, *and* α, *for* $n=1,2,3,\cdots$, *be related by*

$$x_n=\sum_{j=0}^{n-1}\frac{\alpha^{j}}{j!}y_{n-j}. \tag{4.1}$$

Then

$$y_n=\sum_{j=0}^{n-1}(-1)^{j}\frac{\alpha^{j}}{j!}x_{n-j}, \tag{4.1a}$$

and

$$\text{(4.1b)}\qquad \sum_{j=0}^{n-1}\frac{\alpha^j}{j!}\binom{n+r-j-1}{r}y_{n+r-j}=\sum_{m=0}^{r}(-1)^m\frac{\alpha^m}{m!}\binom{n+r-m-1}{r-m}x_{n+r-m}.$$

Proof. The first result (4.1a) is well known. Nielsen obtains (4.1a) by using relations between integrals (see Lewin [17, p. 266]), but mentions that it can be proved directly. An elementary proof follows from the well-known binomial inversion theorem [21, p. 45], namely

$$a_n=\sum_{j=0}^{n}\binom{n}{j}b_{n-j},\qquad b_n=\sum_{j=0}^{n}(-1)^j\binom{n}{j}a_{n-j},$$

by setting $a_0=b_0=0$ and

$$a_k=k!\alpha^{-k}x_k,\qquad b_k=k!\alpha^{-k}y_k.$$

Note that, for fixed p, $L_{n,p}(x)$, $S_{n,p}(x)$, and $M_{n,p}(x)$, $(-1)^pS_{n,p}(-x)$ are, according to (2.15)–(2.18) pairs of functions related by this lemma.

In order to prove (4.1b), Nielsen uses a form of Vandermonde's convolution theorem (see, for example, Riordan [21, pp. 8–10]), namely

$$\text{(4.2)}\qquad \sum_{m=0}^{r}(-1)^m\binom{j}{m}\binom{n+r-m-1}{r-m}=\binom{n-j+r-1}{r},$$

in particular

$$\text{(4.3)}\qquad \sum_{m=0}^{r}(-1)^m\binom{n+k}{m}\binom{n+r-m-1}{r-m}=0,\qquad (0\leqq k\leqq r-1).$$

Denoting the left-hand side of (4.1b) by $A_{n,r}$, we obtain, using (4.2),

$$\text{(4.4)}\qquad A_{n,r}=\sum_{j=0}^{n-1}\frac{\alpha^j}{j!}\sum_{m=0}^{r}(-1)^m\binom{j}{m}\binom{n+r-m-1}{r-m}y_{n+r-j}$$

$$=\sum_{m=0}^{r}(-1)^m\frac{\alpha^m}{m!}\binom{n+r-m-1}{r-m}\sum_{j=0}^{n-m-1}\frac{\alpha^j}{j!}y_{n+r-m-j}.$$

From (4.1), we have

$$\text{(4.5)}\qquad \sum_{j=0}^{n-m-1}\frac{\alpha^j}{j!}y_{n+r-m-j}=x_{n+r-m}-\alpha^{n-m}\sum_{k=0}^{r-1}\frac{\alpha^k}{(n-m+k)!}y_{r-k}.$$

Substituting (4.5) into (4.4), we find as coefficient for y_{r-k} the expression

$$-\frac{\alpha^{n+k}}{(n+k)!}\sum_{m=0}^{r}(-1)^m\binom{n+k}{m}\binom{n+r-m-1}{r-m},\qquad (0\leqq k\leqq r-1)$$

which vanishes by (4.3). The result (4.1b) follows immediately.

5. The transformations $z \to 1-z$ and $z \to 1/z$. Nielsen was able to find many formulae relating the functions $S_{n,p}(z)$, $L_{n,p}(z)$, and $M_{n,p}(z)$ for different arguments. He proved these relations mainly by considering indefinite integrals for $L_{n,p}(z)$ and $M_{n,p}(z)$, for example by writing

$$L_{n,p}(z) = \frac{(-1)^{n+p-1}}{(n-1)!p!} \int \log^{n-1} z \log^p(1-z) \frac{dz}{z} + C,$$

and fixing the constant in the resulting expression by considering special values.

In the following, we shall prove these formulae of Nielsen by using the definitions (2.13) and (2.14) of $L_{n,p}(z)$ and $M_{n,p}(z)$. We start with

5.1. The reflection $z \to 1-z$. We split the range of integration in (2.13), and obtain by partial integration, for $\operatorname{Im} z \neq 0$, using (2.20),

$$\begin{aligned} L_{n,p}(z) &= s_{n,p} + \frac{(-1)^{n+p-1}}{(n-1)!p!} \int_1^z \log^{n-1} t \log^p(1-t) \frac{dt}{t} \\ &= s_{n,p} + \frac{(-1)^{n+p-1}}{n!p!} \log^n z \log^p(1-z) \\ &\quad - \frac{(-1)^{n+p-1}}{(p-1)!n!} \int_0^{1-z} \log^{p-1} t \log^n(1-t) \frac{dt}{t}, \end{aligned}$$

which gives

$$L_{n,p}(z) + L_{p,n}(1-z) = s_{n,p} - \frac{(-1)^{n+p}}{n!p!} \log^n z \log^p(1-z) \tag{5.1}$$

as the reflection formula for the function $L_{n,p}(z)$. Using (2.15) yields

$$S_{n,p}(z) = \sum_{j=0}^{n-1} \frac{\log^j z}{j!} \left\{ s_{n-j,p} - L_{p,n-j}(1-z) - \frac{(-1)^{n+p-j}}{(n-j)!p!} \log^{n-j} z \log^p(1-z) \right\} \tag{5.2}$$

whence, using (2.16) and a well-known property of the binomial coefficients,

$$\begin{aligned} S_{n,p}(z) &= \sum_{j=0}^{n-1} \frac{\log^j z}{j!} \left\{ s_{n-j,p} - \sum_{k=0}^{p-1} \frac{(-1)^k \log^k(1-z)}{k!} S_{p-k,n-j}(1-z) \right\} \\ &\quad + \frac{(-1)^p}{n!p!} \log^n z \log^p(1-z). \end{aligned} \tag{5.3}$$

Equation (5.3) is the reflection formula for the generalized polylogarithms. In the case of polylogarithms ($p=1$), this equation reduces to

$$S_n(z) = \sum_{j=0}^{n-2} \frac{\log^j z}{j!} \{ s_{n-j} - S_{1,n-j-1}(1-z) \} - \frac{1}{(n-1)!} \log^{n-1} z \log(1-z). \tag{5.4}$$

Note that the reflection $z \to 1-z$ for $S_{n,p}(z)$ as expressed by (5.3) remains within the set of generalized polylogarithms (apart from known constants and logarithms). Equation (5.4) shows, however, that it is not possible to represent the reflection for the polylogarithms within the set of these functions if $n \geqq 3$. Thus, for the trilogarithm, one

has

$$S_3(z)=s_3-S_{1,2}(1-z)+\log z\,[s_2-S_2(1-z)]-\tfrac{1}{2}\log^2 z\log(1-z). \tag{5.5}$$

This relation contains the function $S_{1,2}(1-z)$ which is not a polylogarithm. It is interesting to note, however, that the reflection formula provides an expression for the particular generalized polylogarithms $S_{1,p}(z)$ in terms of ordinary polylogarithms, since we have

$$S_{1,p}(z)=\frac{(-1)^p}{p!}\int_0^1 \log^p(1-zt)\frac{dt}{t}=s_{p+1}+\frac{(-1)^p}{p!}\log z\log^p(1-z) - \sum_{k=0}^{p-1}(-1)^k\log^k\frac{(1-z)}{k!}S_{p-k+1}(1-z). \tag{5.6}$$

We now consider the case $z=x$ real. By setting $z=x+i\varepsilon$ it follows from (3.2), (3.3) and (1.4) that, for $x<0$, (5.1) remains unchanged, while for $x>1$, both sides must be replaced by their complex conjugate. From this and the fact that $L_{n,p}(x)$ is real for $0\leqq x\leqq 1$, we see that (5.1) remains valid, as it stands, for arbitrary real $z=x$. By a similar argument, this is also true for (5.3) relating $S_{n,p}(z)$ to $S_{\nu,\rho}(1-z)$, and for the special case (5.4). Therefore, the restriction $\operatorname{Im} z\neq 0$ made above is no longer necessary.

5.2. The inversion $z\to 1/z$. This case is more complicated. We follow Nielsen in first establishing an inversion formula for $M_{n,p}(z)$ but instead of considering indefinite integrals for $M_{n,p}(z)$ and $M_{n,p}(1/z)$ as he does, we start with the definition integral (2.14), which gives for $\operatorname{Im} z\neq 0$, using (2.20),

$$\begin{aligned}
M_{n,p}(z)&=\sigma_{n,p}+\frac{(-1)^{n-1}}{(n-1)!p!}\int_1^z \log^{n-1}t\log^p(1+t)\frac{dt}{t}\\
&=\sigma_{n,p}-\frac{1}{(n-1)!p!}\int_1^{1/z}\log^{n-1}t\,[\log(1+t)-\log t]^p\frac{dt}{t}\\
&=\sigma_{n,p}-\frac{1}{(n-1)!p!}\sum_{k=0}^{p-1}(-1)^k\binom{p}{k}\int_1^{1/z}\log^{n+k-1}t\log^{p-k}(1+t)\frac{dt}{t}\\
&\quad-\frac{(-1)^p}{(n-1)!p!}\int_1^{1/z}\log^{n+p-1}t\,\frac{dt}{t}\\
&=\sigma_{n,p}-(-1)^n\sum_{k=0}^{p-1}\binom{n+k-1}{k}\sigma_{n+k,p-k}\\
&\quad+(-1)^n\sum_{k=0}^{p-1}\binom{n+k-1}{k}M_{n+k,p-k}\left(\frac{1}{z}\right)-\frac{(-1)^n}{(n+p)(n-1)!p!}\log^{n+p}z.
\end{aligned}$$

Denoting the constant part by $C_{n,p}$, we have

$$M_{n,p}(z)=C_{n,p}+(-1)^n\sum_{k=0}^{p-1}\binom{n+k-1}{k}M_{n+k,p-k}\left(\frac{1}{z}\right)-\frac{(-1)^n}{(n+p)(n-1)!p!}\log^{n+p}z \tag{5.7}$$

$$(\operatorname{Im} z\neq 0,\ \text{or}\ \operatorname{Im} z=0\ \text{and}\ \operatorname{Re} z\geqq 0).$$

Formula (5.7) is also valid for real $z=x\geqq 0$, in which case it is real. For $x<0$, we set $z=x+i\varepsilon$ and obtain with the help of (1.4) and (3.5) the relation

$$M_{n,p}(x)=C_{n,p}+(-1)^n\sum_{k=0}^{p-1}\binom{n+k-1}{k}M^*_{n+k,p-k}\left(\frac{1}{x}\right) - \frac{(-1)^n}{(n+p)(n-1)!p!}\log^{n+p}x, \qquad (x \text{ real}). \tag{5.8}$$

This equation enables us to express the constants (where the sum is zero for $p=1$)

$$C_{n,p}=(1-(-1)^n)\sigma_{n,p}-(-1)^n\sum_{k=1}^{p-1}\binom{n+k-1}{k}\sigma_{n+k,p-k} \tag{5.9}$$

in terms of known constants. This will be discussed further in §7.

In order to establish the inversion formula for the generalized polylogarithms $S_{n,p}(z)$, we start from (5.7) and write, using (2.17)

(5.10)

$$(-1)^pS_{n,p}(-z)=\sum_{j=0}^{n-1}\frac{\log^j z}{j!}(-1)^{n-j}\sum_{k=0}^{p-1}\binom{n-j+k-1}{k}M_{n-j+k,p-k}\left(\frac{1}{z}\right) - \frac{1}{p!}\log^{n+p}z\sum_{j=0}^{n-1}\frac{(-1)^{n-j}}{(n-j+p)(n-j-1)!j!}+\sum_{j=0}^{n-1}\frac{\log^j z}{j!}C_{n-j,p}.$$

The coefficient of $\log^{n+p}z$ can be written as [8, No. 3.191 3]

(5.11)

$$-\frac{1}{p!}\sum_{j=0}^{n-1}\frac{(-1)^{n-j}}{(n-j+p)(n-j-1)!j!}=\frac{1}{p!(n-1)!}\int_0^1\sum_{j=0}^{n-1}(-1)^{n-j}\binom{n-1}{j}u^{n-j+p-1}du =\frac{1}{p!(n-1)!}\int_0^1 u^p(1-u)^{n-1}du=\frac{1}{(n+p)!}.$$

We now apply Lemma (4.1b) and write, remembering that $M_{n,p}(z)$ and $(-1)^pS_{n,p}(-z)$ are related through (4.1),

$$\sum_{j=0}^{n-1}(-1)^j\frac{\log^j z}{j!}\binom{n-j+k-1}{k}M_{n-j+k,p-k}\left(\frac{1}{z}\right) =(-1)^{p-k}\sum_{m=0}^{k}\frac{\log^m z}{m!}\binom{n+k-m-1}{k-m}S_{n+k-m,p-k}\left(-\frac{1}{z}\right)$$

whence, replacing z by $-z$ in (5.10)

$$S_{n,p}(z)=(-1)^n\sum_{k=0}^{p-1}(-1)^k\sum_{m=0}^{k}\frac{\log^m(-z)}{m!}\binom{n+k-m-1}{k-m}S_{n+k-m,p-k}\left(\frac{1}{z}\right) +(-1)^p\left\{\sum_{j=0}^{n-1}\frac{\log^j(-z)}{j!}C_{n-j,p}+\frac{1}{(n+p)!}\log^{n+p}(-z)\right\}. \tag{5.12}$$

It is easy to see that (5.12) is also valid for real $z=x$ if $x\leqq 0$. In this case it contains only real functions. For $x>0$, we proceed analogously to the case of $M_{n,p}(z)$, set $z=x+i\varepsilon$ and obtain, using (1.4) and (3.1),

(5.13)

$$S_{n,p}(x)=(-1)^n\sum_{k=0}^{p-1}(-1)^k\sum_{m=0}^{k}\frac{\log^m(-x)}{m!}\binom{n+k-m-1}{k-m}S^*_{n+k-m,p-k}\left(\frac{1}{x}\right)$$

$$+(-1)^p\left\{\sum_{j=0}^{n-1}\frac{\log^j(-x)}{j!}C_{n-j,p}+\frac{1}{(n+p)!}\log^{n+p}(-x)\right\}\qquad (x \text{ real}).$$

Note that the imaginary parts cancel for $0\leqq x\leqq 1$. Further, since $S_{\nu,\rho}(1/x)$ is real for $x\geqq 1$, it follows that the imaginary part of $S_{n,p}(x)$ for $x\geqq 1$ comes only from $\log(-x)$. Using (1.4) in (5.13) and taking the imaginary part (with empty sums equal to zero) gives[1]

(5.14)

$$\frac{1}{\pi}\operatorname{Im}S_{n,p}(x)=(-1)^n\sum_{k=1}^{p-1}(-1)^k\sum_{m=1}^{k}\binom{n+k-m-1}{k-m}S_{n+k-m,p-k}\left(\frac{1}{x}\right)$$

$$\times\sum_{j=0}^{[(m-1)/2]}(-1)^j\frac{\pi^{2j}}{(2j+1)!(m-2j-1)!}\log^{m-2j-1}x$$

$$+(-1)^p\sum_{m=1}^{n-1}C_{n-m,p}\sum_{j=0}^{[(m-1)/2]}(-1)^j\frac{\pi^{2j}}{(2j+1)!(m-2j-1)!}\log^{m-2j-1}x$$

$$+(-1)^p\sum_{j=0}^{[(n+p-1)/2]}(-1)^j\frac{\pi^{2j}}{(2j+1)!(n+p-2j-1)!}\log^{n+p-2j-1}x.$$

Equation (5.14) shows that, for $p>1$, $\operatorname{Im}S_{n,p}(x)$ is expressible, apart from logarithms and (known) constants, in terms of $S_{\nu,\rho}(\xi)$ $(n\leqq\nu\leqq n+p-2, 1\leqq\rho\leqq p-1)$ in the interval $0<\xi\leqq 1$.

For the polylogarithms, we obtain as a special case from (5.12), (5.9) and (2.22)

(5.15) $$S_n(z)+(-1)^nS_n\left(\frac{1}{z}\right)$$

$$=-\frac{1}{n!}\log^n(-z)-\sum_{j=0}^{n-2}\frac{1}{j!}\left(1+(-1)^{n-j}\right)\left(1-2^{1-n+j}\right)s_{n-j}\log^j(-z),$$

where for real $z=x$, $S_n(1/x)$ has to be replaced by $S_n^*(1/x)$.

As in the case of reflection, the inversion (5.12) has the remarkable property of remaining within the set of generalized polylogarithms. In contrast to the case of reflection, the inversion (5.15) for the polylogarithms also remains within the set of these functions.

---

[1] Note that this formula differs by a factor -1 from the expression given in [12], owing to the fact that, for real $z=x$, (5.12) had been erroneously used instead of (5.13), and a convention different from (1.4) had been introduced in order to overcome resulting contradictions. Because of this, the function computed by the Algol program for $S_{n,p}(x)$ in [12] is in fact $S^*_{n,p}(x)$ according to our present definition. For $n=p=1$, formula (5.14) agrees with Lewin [17, p. 2].

6. Bilinear transformations of the argument. If we consider the reflection $P_1(z)=1-z$ and the inversion $P_2(z)=1/z$ as belonging to the group of bilinear transformations

$$P(z)=\frac{\alpha z+\beta}{\gamma z+\delta}$$

with real coefficients and $\alpha\delta-\beta\gamma\neq 0$, we see that they generate a subgroup

$$P_j(z)=z,\quad 1-z,\quad \frac{1}{z},\quad \frac{1}{1-z},\quad \frac{z-1}{z},\quad \frac{z}{z-1}$$

($j=0,1,\cdots,5$). By repeated use of (5.3) and (5.12) it is then possible to find formulae for

$$S_{n,p}\left(\frac{1}{1-z}\right),\quad S_{n,p}\left(\frac{z-1}{z}\right),\quad S_{n,p}\left(\frac{z}{z-1}\right)$$

which will contain, apart from logarithms and known constants, only generalized polylogarithms. These formulae become soon very complicated, and we leave a more systematic investigation to further research. By the above procedure one can, in principle, find a (complicated) expression for $S_{2,2}(z)$ in terms of polylogarithms. Lewin [17, p. 204] has given a formula from which such an expression can be derived.

Nielsen also discusses the following bilinear transformations of the argument of the function $L_{n,p}(z)$.

6.1. The transformation $z\to 1/(1+z)$. From the definition integral (2.13) we find, for $\operatorname{Im} z\neq 0$,

(6.1)

$$\begin{aligned}L_{n,p}\left(\frac{1}{1+z}\right)&=\frac{(-1)^{n+p-1}}{(n-1)!p!}\int_0^{1/(1+z)}\log^{n-1}t\log^p(1-t)\frac{dt}{t}\\&=s_{n,p}-\frac{1}{(n-1)!p!}\int_0^z\log^{n-1}(1+t)[\log(1+t)-\log t]^p\frac{dt}{1+t}\\&=s_{n,p}-\frac{1}{(n-1)!p!}\sum_{k=0}^{p}(-1)^k\binom{p}{k}\int_0^z\log^{n+p-k-1}(1+t)\log^k t\frac{dt}{1+t}.\end{aligned}$$

Using (2.14), and taking into account that

$$\frac{\binom{p}{k}}{(n-1)!p!(n+p-k)}=\frac{\binom{n+p-k-1}{n-1}\binom{n+p}{k}}{(n+p)!},$$

yields, after integration by parts,

(6.2)

$$\begin{aligned}L_{n,p}\left(\frac{1}{1+z}\right)&=s_{n,p}-\frac{1}{(n+p)!}\sum_{k=0}^{p}(-1)^k\binom{n+p-k-1}{n-1}\binom{n+p}{k}\log^k z\log^{n+p-k}(1-z)\\&\quad-\sum_{k=1}^{p}\binom{n+p-k-1}{n-1}M_{k,n+p-k}(z)\qquad(\operatorname{Im} z\neq 0,\text{ or }\operatorname{Im} z=0\text{ and }\operatorname{Re} z\geqq -1).\end{aligned}$$

Both sides of formula (6.2) are real for $z=x\geqq 0$. Further, it follows from (1.4) and (3.3) that (6.2) remains valid for real $z=x$ if $-1\leqq x\leqq 0$. For $z=x<-1$, however, we see from (3.2) that $L_{n,p}(1/(1+z))$ has to be replaced by $L^*_{n,p}(1/(1+z))$.

6.2. The transformation $z\to z/(1+z)$. In this case, we have from (2.13) for $\operatorname{Im} z\neq 0$ and $n>1$

$$L_{n,p}\left(\frac{z}{1+z}\right)=\frac{(-1)^{n+p-1}}{(n-1)!p!}\int_0^{z/(1+z)}\log^{n-1}t\log^p(1-t)\frac{dt}{t}$$

$$=\frac{1}{(n-1)!p!}\left\{\int_0^z[\log(1+t)-\log t]^{n-1}\log^p(1+t)\frac{dt}{t}\right.$$

$$\left.-\int_0^z[\log(1+t)-\log t]^{n-1}\log^p(1+t)\frac{dt}{1+t}\right\}.$$

Hence by comparison with (6.1),

$$L_{n,p}\left(\frac{z}{1+z}\right)=L_{p+1,n-1}\left(\frac{1}{1+z}\right)-s_{p+1,n-1}$$

$$+\frac{1}{(n-1)!p!}\sum_{k=0}^{n-1}(-1)^k\binom{n-1}{k}\int_0^z\log^{n+p-k-1}(1+t)\log^k t\frac{dt}{t}$$

and, using (2.14) and (2.21),

(6.3)

$$L_{n,p}\left(\frac{z}{1+z}\right)-L_{p+1,n-1}\left(\frac{1}{1+z}\right)=-s_{n-1,p+1}+\sum_{k=0}^{n-1}\binom{n+p-k-1}{p}M_{k+1,n+p-k-1}(z)$$

$$(\operatorname{Im} z\neq 0,\text{ or }\operatorname{Im} z=0\text{ and }\operatorname{Re} z\geqq -1).$$

For $n=1$, we have directly from (2.13):

(6.4)
$$L_{1,p}\left(\frac{z}{1+z}\right)=\frac{1}{p!}\left\{\int_0^z\log^p(1+t)\frac{dt}{t}-\int_0^z\log^p(1+t)\frac{dt}{1+t}\right\}$$

$$=M_{1,p}(z)-\frac{1}{(p+1)!}\log^{p+1}(1+z)$$

or, by (2.19),

(6.5)
$$S_{1,p}\left(\frac{z}{1+z}\right)=(-1)^pS_{1,p}(-z)-\frac{1}{(p+1)!}\log^{p+1}(1+z)$$

$$(\operatorname{Im} z\neq 0,\text{ or }\operatorname{Im} z=0\text{ and }\operatorname{Re} z\geqq -1).$$

As in the case of formula (6.2), we see by setting $z=x+i\varepsilon$ and using (1.4), (3.2) and (3.3) that the left-hand side of (6.2) must be replaced, for real $z=x<-1$, by

$$L^*_{n,p}\left(\frac{z}{1+z}\right)-L^*_{p+1,n-1}\left(\frac{1}{1+z}\right).$$

Since $L_{1,p}(z/(1+z))=S_{1,p}(z/(1+z))$ is real for $x\geqq -1$, the formula obtained by replacing the left-hand sides of (6.4) and (6.5) by their conjugates are valid for all real $z=x$.

6.3. The transformation $z \to (a+z)/(b+z)$. This transformation is a generalization of the transformation $z \to z/(1+z)$. The formula obtained by Nielsen [20, §13(7)] is incorrect, and, as we shall see in §9, it was this error which led him to wrong conclusions on the nature of the special values $\sigma_{n,p} = (-1)^p S_{n,p}(-1)$.

We have from the definition integral (2.13) for $a \neq b$, $\operatorname{Im} z \neq 0$, and $n > 1$,

$$L_{n,p}\left(\frac{a+z}{b+z}\right) = \frac{(-1)^{n+p-1}}{(n-1)!p!}\int_0^{(a+z)/(b+z)} \log^{n-1} t \log^p(1-t)\frac{dt}{t}$$

$$= s_{n,p} + \frac{(-1)^{n+p}}{(n-1)!p!}(b-a)\int_z^\infty \log^{n-1}\left(\frac{a+t}{b+t}\right)\log^p\left(\frac{b-a}{b+t}\right)\frac{dt}{(a+t)(b+t)}$$

$$= s_{n,p} + \frac{(-1)^{n+p}}{(n-1)!p!}\left\{\int_z^\infty \log^{n-1}\left(\frac{a+t}{b+t}\right)\log^p\left(\frac{b-a}{b+t}\right)\frac{dt}{a+t}\right.$$

$$\left. - \int_z^\infty \log^{n-1}\left(\frac{a+t}{b+t}\right)\log^p\left(\frac{b-a}{b+t}\right)\frac{dt}{b+t}\right\}.$$

Denoting the integrals by J_1 and J_2, and substituting $\tau = (b-a)/(a+t)$ and $\tau = (b-a)/(b+t)$ in J_1 and J_2, respectively, we obtain, using (2.14),

$$J_1 = (-1)^{n+p-1}\int_0^{(b-a)/(a+z)} \log^{n-1}(1+\tau)\log^p\left(\frac{1+\tau}{\tau}\right)\frac{d\tau}{\tau}$$

$$= (-1)^{n+p-1}\sum_{k=0}^{p}(-1)^k\binom{p}{k}\int_0^{(b-a)/(a+z)} \log^{n+p-k-1}(1+\tau)\log^k\tau\frac{d\tau}{\tau}$$

$$= (-1)^{n+p-1}p!(n-1)!\sum_{k=0}^{p}\binom{n+p-k-1}{n-1}M_{k+1,n+p-k-1}\left(\frac{b-a}{a+z}\right).$$

Similarly, by (2.13),

$$J_2 = \int_0^{(b-a)/(b+z)} \log^{n-1}(1-\tau)\log^p\tau\frac{d\tau}{\tau}$$

$$= (-1)^{n+p-1}p!(n-1)!L_{p+1,n-1}\left(\frac{b-a}{b+z}\right).$$

Therefore

(6.6)

$$L_{n,p}\left(\frac{a+z}{b+z}\right) = s_{n,p} - \sum_{k=0}^{p}\binom{n+p-k-1}{n-1}M_{k+1,n+p-k-1}\left(\frac{b-a}{a+z}\right) + L_{p+1,n-1}\left(\frac{b-a}{b+z}\right)$$

$$(\operatorname{Im} z \neq 0).$$

This formula corrects Nielsen's formula [20, §13(7)], in which

$$\text{the term} \quad L_{p+1,n-1}\left(\frac{b-a}{b+z}\right) \quad \text{reads} \quad L_{n,p}\left(\frac{b-a}{b+z}\right).$$

For real $z = x$, formula (6.6) has to be adjusted in accordance with the results of §3. This will depend on the values of a and b.

Comparing formula (6.3) with (6.6) for $a=0$ and $b=1$, we obtain a relation between $M_{\nu,\rho}(z)$ and $M_{\nu,\rho}(1/z)$, namely $(n>1)$

$$(6.7)\qquad \sum_{k=0}^{n-1}\binom{n+p-k-1}{p}M_{k+1,n+p-k-1}(z)$$
$$+\sum_{k=0}^{p}\binom{n+p-k-1}{n-1}M_{k+1,n+p-k-1}\left(\frac{1}{z}\right)=s_{n,p}+s_{n-1,p+1}.$$

$(\operatorname{Im} z\neq 0$, or $\operatorname{Im} z=0$ and $\operatorname{Re} z\geqq 0)$.

For real $z=x<0$, $M_{k+1,n+p-k-1}(1/z)$ has to be replaced by $M^{*}_{k+1,n+p-k-1}(1/z)$.

6.4. A sum containing $L_{n,p}(z/(1+z))$. Nielsen closes his section on bilinear transformations of the argument by considering a sum containing $L_{\nu,\rho}(z/(1+z))$. This sum can be expressed in terms of (ordinary) polylogarithms. To avoid divergent integrals, we introduce the functions $(0<\varepsilon<1)$

$$(6.8)\qquad \left.\begin{matrix}L^{\varepsilon}_{n,p}(z)\\(-1)^{p}M^{\varepsilon}_{n,p}(z)\end{matrix}\right\}=\frac{(-1)^{n+p-1}}{(n-1)!p!}\int_{\varepsilon}^{z}\log^{n-1}t\log^{p}(1\mp t)\frac{dt}{t}$$

and write

$$(-1)^{j}k!L^{\varepsilon}_{k-j+1,j}(\zeta)=(-1)^{j}k!\frac{(-1)^{k}}{(k-j)!j!}\int_{\varepsilon}^{\zeta}\log^{k-j}t\log^{j}(1-t)\frac{dt}{t}.$$

Thus, by summation,

$$\sum_{j=0}^{k}(-1)^{j}L^{\varepsilon}_{k-j+1,j}(\zeta)=\frac{(-1)^{k}}{k!}\sum_{j=0}^{k}(-1)^{j}\binom{k}{j}\int_{\varepsilon}^{\zeta}\log^{k-j}t\log^{j}(1-t)\frac{dt}{t}$$
$$=\frac{(-1)^{k}}{k!}\int_{\varepsilon}^{\zeta}\log^{k}\left(\frac{t}{1-t}\right)\frac{dt}{t}.$$

Setting $\zeta=z/(1+z)$, and making the substitution $\tau=t/(1-t)$, we obtain, with $\varepsilon'=\varepsilon/(1-\varepsilon)$,

$$\sum_{j=0}^{k}(-1)^{j}L^{\varepsilon}_{k-j+1,j}\left(\frac{z}{1+z}\right)=\frac{(-1)^{k}}{k!}\int_{\varepsilon'}^{z}\log^{k}\tau\frac{d\tau}{\tau(1+\tau)}$$
$$=\frac{(-1)^{k}}{(k+1)!}\left[\log^{k+1}z-\log^{k+1}\varepsilon'\right]-\frac{(-1)^{k}}{k!}\int_{\varepsilon'}^{z}\log^{k}\tau\frac{d\tau}{1+\tau}.$$

By partial integration, using (6.8)

$$\int_{\varepsilon'}^{z}\log^{k}\tau\frac{d\tau}{1+\tau}=\log^{k}z\log(1+z)-\log^{k}\varepsilon'\log(1+\varepsilon')+(-1)^{k}k!M^{\varepsilon'}_{k,1}(z).$$

Further, from (6.8),

$$L^{\varepsilon}_{k+1,0}\left(\frac{z}{1+z}\right)=\frac{(-1)^{k}}{(k+1)!}\left[\log^{k+1}\left(\frac{z}{1+z}\right)-\log^{k+1}\varepsilon\right].$$

Collecting these results, we see that the divergent parts cancel for $\varepsilon \to 0$, and we obtain

$$\sum_{j=1}^{k} (-1)^j L_{k-j+1,j}\left(\frac{z}{1+z}\right) = \frac{(-1)^k}{(k+1)!}\left[\log^{k+1} z - \log^{k+1}\left(\frac{z}{1+z}\right)\right] - \frac{(-1)^k}{k!}\log^k z \log(1+z) - M_{k,1}(z) \tag{6.9}$$

or, using (2.18) and replacing k by $n-1$,

$$\sum_{m=0}^{n-2} (-1)^m L_{n-m-1,m+1}\left(\frac{z}{1+z}\right) = \frac{(-1)^n}{n!}\left[\log^n z - \log^n\left(\frac{z}{1+z}\right)\right] - \sum_{j=0}^{n-1} \frac{(-1)^j \log^j z}{j!} S_{n-j}(-z). \tag{6.10}$$

As in the case of formula (6.2), the left-hand side of (6.10) has to be replaced by its conjugate for real $z = x < -1$.

7. The constants $C_{n,p}$. We now return to the constants $C_{n,p}$ defined by (5.9). Because no relation seems to be known which expresses $\sigma_{n,p} = (-1)^p S_{n,p}(-1)$ for arbitrary $n \geqq 1$, $p \geqq 1$ in terms of known constants, Nielsen did not, apart from giving a formula for $C_{2n-1,2}$ in terms of $\zeta(q)$, investigate relation (5.9) further. For some small values of n or p, however, expressions for $C_{n,p}$ are known [12] in addition to $C_{2n-1,2}$, namely

$$\begin{aligned} &C_{1,p} = \zeta(p+1), \qquad (p = 1,2,3,4), \\ &C_{2n,1} = 0, \qquad (n \geqq 1), \\ &C_{3,1} = -C_{2,2} = \frac{7}{4}\zeta(4) = \frac{7}{360}\pi^4, \\ &C_{2,3} = -C_{3,2} = -\zeta(2)\zeta(3) - \zeta(5). \end{aligned} \tag{7.1}$$

The question arises whether it is possible to express $C_{n,p}$ for all n and p, in terms of known constants. This is indeed the case, as is shown by the following

THEOREM 1. *Let δ_{ij} be the Kronecker symbol; let $\varepsilon_m = 1$ if m is even, $\varepsilon_m = 0$ if m is odd, and let $s_{\nu,\rho} = S_{\nu,\rho}(1)$. Then, for integer n,p $(n \geqq 1, p \geqq 1)$*

$$C_{n,p} = (-1)^{n+p-1} \sum_{k=0}^{p-1} (-1)^k \left(1 - (-1)^n \delta_{k0}\right) \binom{n+k-1}{k} \times \sum_{j=0}^{[(n+k-1)/2]} (-1)^j \frac{\pi^{2j}}{(2j)!} s_{n+k-2j,p-k} + (-1)^{[(n+p)/2]+n} \varepsilon_{n+p} \frac{\pi^{n+p}}{(n+p)(n-1)!p!}. \tag{7.2}$$

Proof. We set $x=-1$ in (5.8) and obtain, using (1.4),

$$C_{n,p}=M_{n,p}(-1)-(-1)^n\sum_{k=0}^{p-1}\binom{n+k-1}{k}M^*_{n+k,p-k}(-1)+(-1)^n\frac{(i\pi)^{n+p}}{(n+p)(n-1)!p!}.$$

From (2.18), we have

$$M_{n,p}(-1)=(-1)^p\sum_{j=0}^{n-1}\frac{(-i\pi)^j}{j!}s_{n-j,p}. \tag{7.3}$$

Therefore

(7.4)

$$C_{n,p}=(-1)^{n+p-1}\Bigg\{\sum_{k=0}^{p-1}(-1)^k\binom{n+k-1}{k}\sum_{j=0}^{n+k-1}\frac{(i\pi)^j}{j!}s_{n+k-j,p-k} -(-1)^n\sum_{j=0}^{n-1}(-1)^j\frac{(i\pi)^j}{j!}s_{n-j,p}\Bigg\} +(-1)^n\frac{(i\pi)^{n+p}}{(n+p)(n-1)!p!}.$$

On taking the real part of (7.4), Theorem 1 follows. In addition, we obtain from the imaginary part a relation between the $s_{\nu,\rho}$, namely

$$\sum_{k=0}^{p-1}(-1)^k\left(1+(-1)^n\delta_{k0}\right)\binom{n+k-1}{k}\sum_{j=0}^{[(n+k-2)/2]}(-1)^j\frac{\pi^{2j+1}}{(2j+1)!}s_{n+k-2j-1,p-k} +(-1)^{[(n+p-1)/2]+p}\left(\varepsilon_{n+p}-1\right)\frac{\pi^{n+p}}{(n+p)(n-1)!p!}=0.$$

It has been shown in [12], [13], [20] that $s_{\nu,\rho}$ can be expressed by a homogeneous polynomial in $s_q=\zeta(q)$ $(2\leqq q\leqq\nu+\rho)$, with rational coefficients, where $\zeta(q)$ is the Riemann zeta function. This polynomial is of degree $\nu+\rho$, if $\zeta(q)$ is defined to be of degree q and if the degree of a product is the sum of the degree of its factors. Since even powers of π can, through [8, No. 9.616]

$$\pi^{2m}=\frac{(2m)!\zeta(2m)}{2^{2m-1}|B_{2m}|}$$

be expressed in terms of $\zeta(q)$ and the Bernoulli numbers B_{2m}, one has the remarkable fact that $C_{n,p}$ can be expressed by a polynomial in $\zeta(q)$ $(2\leqq q\leqq n+p)$ which is of the same type as the polynomial for $s_{n,p}$.

A short table of $C_{n,p}$ $(1\leqq n\leqq 5, 1\leqq p\leqq 5)$ is given in Table 1. This table has been computed by REDUCE [10] from the expressions for $s_{r,m}$ given in [13]. For the special case $n=1$, we find from (5.9) and Theorem 3 in §9 that

$$C_{1,p}=2\sigma_{1,p}+\sum_{k=1}^{p-1}\sigma_{k+1,p-k}=\zeta(p+1), \qquad (p\geqq 1), \tag{7.5}$$

which gives, with (7.2), another relation between the $s_{\nu,\rho}$.

TABLE 1

$$C_{1,1}=\zeta(2)=\frac{\pi^2}{6} \qquad C_{2,1}=0$$

$$C_{1,2}=\zeta(3) \qquad C_{2,2}=-6\zeta^2(2)+\frac{53}{4}\zeta(4)=-\frac{7}{360}\pi^4$$

$$C_{1,3}=\zeta(4)=\frac{\pi^4}{90} \qquad C_{2,3}=-\zeta(2)\zeta(3)-\zeta(5)$$

$$C_{1,4}=\zeta(5) \qquad C_{2,4}=-\frac{25}{6}\zeta^3(2)+\frac{21}{2}\zeta(2)\zeta(4)-\frac{1}{2}\zeta^2(3)-\frac{163}{48}\zeta(6)$$

$$C_{1,5}=\zeta(6)=\frac{\pi^6}{945} \qquad C_{2,5}=-\frac{5}{2}\zeta^2(2)\zeta(3)-\zeta(2)\zeta(5)+\frac{21}{4}\zeta(3)\zeta(4)-2\zeta(7)$$

$$C_{3,1}=6\zeta^2(2)-\frac{53}{4}\zeta(4)=\frac{7}{360}\pi^4$$

$$C_{3,2}=\zeta(2)\zeta(3)+\zeta(5)$$

$$C_{3,3}=\frac{25}{6}\zeta^3(2)-12\zeta(2)\zeta(4)+\frac{1}{2}\zeta^2(3)+\frac{191}{24}\zeta(6)$$

$$C_{3,4}=\frac{17}{2}\zeta^2(2)\zeta(3)+2\zeta(2)\zeta(5)-\frac{37}{2}\zeta(3)\zeta(4)+3\zeta(7)$$

$$C_{3,5}=\frac{23}{8}\zeta^4(2)-\frac{17}{2}\zeta^2(2)\zeta(4)+\frac{1}{2}\zeta(2)\zeta^3(3)+\frac{531}{16}\zeta(2)\zeta(6)+2\zeta(3)\zeta(5)-\frac{175}{8}\zeta^2(4)-\frac{939}{64}\zeta(8)$$

$$C_{4,1}=0$$

$$C_{4,2}=3\zeta(2)\zeta(4)-\frac{73}{8}\zeta(6)=-\frac{31}{7560}\pi^6$$

$$C_{4,3}=-12\zeta^2(2)\zeta(3)-2\zeta(2)\zeta(5)+\frac{53}{2}\zeta(3)\zeta(4)-2\zeta(7)$$

$$C_{4,4}=-2\zeta^4(2)+\frac{35}{4}\zeta^2(2)\zeta(4)-\zeta(2)\zeta^2(3)-\frac{205}{4}\zeta(2)\zeta(6)-2\zeta(3)\zeta(5)+\frac{131}{4}\zeta^2(4)+\frac{1357}{64}\zeta(8)$$

$$C_{4,5}=-\frac{25}{3}\zeta^3(2)\zeta(3)-23\zeta^2(2)\zeta(5)+\frac{45}{2}\zeta(2)\zeta(3)\zeta(4)-5\zeta(2)\zeta(7)-\frac{1}{3}\zeta^3(3)-\frac{545}{48}\zeta(3)\zeta(6)+\frac{201}{4}\zeta(4)\zeta(5)-\frac{23}{3}\zeta(9)$$

$$C_{5,1}=-\frac{3}{2}\zeta(2)\zeta(4)+\frac{73}{16}\zeta(6)=\frac{31}{15120}\pi^6$$

$$C_{5,2}=6\zeta^2(2)\zeta(3)+\zeta(2)\zeta(5)-\frac{53}{4}\zeta(4)+\zeta(7)$$

$$C_{5,3}=\zeta^4(2)-\frac{35}{8}\zeta^2(2)\zeta(4)+\frac{1}{2}\zeta(2)\zeta^2(3)+\frac{755}{16}\zeta(2)\zeta(6)+\zeta(3)\zeta(5)-\frac{281}{8}\zeta^2(4)-\frac{1261}{64}\zeta(8)$$

$$C_{5,4}=\frac{25}{6}\zeta^3(2)\zeta(3)+\frac{53}{2}\zeta^2(2)\zeta(5)-15\zeta(2)\zeta(3)\zeta(4)+5\zeta(2)\zeta(7)+\frac{1}{6}\zeta^3(3)+\frac{205}{12}\zeta(3)\zeta(6)-\frac{233}{4}\zeta(4)\zeta(5)+\frac{19}{3}\zeta(9)$$

$$C_{5,5}=\frac{73}{120}\zeta^5(2)+\frac{109}{24}\zeta^3(2)\zeta(4)+\frac{41}{4}\zeta^2(2)\zeta^2(3)+\frac{4405}{96}\zeta^2(2)\zeta(6)+4\zeta(2)\zeta(3)\zeta(5)-62\zeta(2)\zeta^2(4)$$
$$+\frac{1659}{32}\zeta(2)\zeta(8)-\frac{45}{2}\zeta^2(3)\zeta(4)+5\zeta(3)\zeta(7)-\frac{3385}{96}\zeta(4)\zeta(6)+\frac{5}{2}\zeta^2(5)-\frac{22177}{640}\zeta(10)$$

8. Some integrals involving $S_{n,p}(z)$. We now consider some integrals containing $S_{n,p}(z)$ in the integrand which were discussed by Nielsen. From the differentiation rule (2.11) for $S_{n,p}(x)$ we see immediately that

$$\int S_{n-1,p}(z)S_{n,p}(z)\frac{dz}{z}=\frac{1}{2}S_{n,p}^2(z), \qquad (n\geqq 2). \tag{8.1}$$

By repeated partial integration we obtain for $k<n-1$, (note the misprints in Nielsen's formulae [20, p. 190])

$$\int S_{n-1,p}(z)S_{n,p}(z)\frac{dz}{z}=\sum_{j=1}^{k}(-1)^{j-1}S_{n-j,p}(z)S_{n+j,p}(z)$$
$$+(-1)^k\int S_{n-k-1,p}(z)S_{n+k,p}(z)\frac{dz}{z}.$$

Therefore

(8.2)

$$(-1)^k \int S_{n-k-1,p}(z) S_{n+k,p}(z) \frac{dz}{z} = \frac{1}{2} S_{n,p}^2(z) + \sum_{j=1}^{k} (-1)^j S_{n-j,p}(z) S_{n+j,p}(z)$$

$$= \frac{1}{2}(-1)^k \sum_{j=0}^{2k} (-1)^j S_{n-k+j,p}(z) S_{n+k-j,p}(z).$$

Using the definitions (2.6) of $S_0(z)$ and $S_1(z)$, and rule (2.12), we obtain, by repeated partial integration, for $\alpha \neq 1$ and $|z|<1$,

$$\int z^\alpha S_n(z)\,dz = z^{\alpha+1} \sum_{j=0}^{n-1} (-1)^j \frac{S_{n-j}(z)}{(\alpha+1)^{j+1}} + \frac{(-1)^n}{(\alpha+1)^n} \int \frac{z^{\alpha+1}}{1-z}\,dz \tag{8.3}$$

$$= z^{\alpha+1} \sum_{j=0}^{n-2} \frac{(-1)^j S_{n-j}(z)}{(\alpha+1)^{j+1}} + \frac{(-1)^n}{(\alpha+1)^n}(z^{\alpha+1}-1)\log(1-z)$$

$$+ \frac{(-1)^{n-1}}{(\alpha+1)^n} \int \frac{1-z^{\alpha+1}}{1-z}\,dz,$$

and hence, using [8, No. 8.361 7], we obtain the definite integral

(8.4)

$$\int_0^1 x^{\alpha-1} S_n(x)\,dx = \sum_{j=0}^{n-2} \frac{(-1)^j s_{n-j}}{\alpha^{j+1}} + \frac{(-1)^{n-1}}{\alpha^n}[\psi(\alpha+1)+\gamma], \qquad (\mathrm{Re}\,\alpha > -1),$$

where $\psi(x)$ is the logarithmic derivative of the gamma function and γ is Euler's constant. In particular, for $\alpha = m > 0$ an integer,

$$\int_0^1 x^{m-1} S_n(x)\,dx = \sum_{j=0}^{n-2} \frac{(-1)^j s_{n-j}}{m^{j+1}} + \frac{(-1)^{n-1}}{m^n} \sum_{k=1}^{m} \frac{1}{k}. \tag{8.5}$$

Lewin [17, p. 308] gives (8.4) and (8.5) for $n=2$.

Equation (8.5), together with the power series expansion (2.5), allows us to compute

$$\phi(x) = \int_0^1 S_q(tx) S_n(t) \frac{dt}{t} = \sum_{m=1}^{\infty} \frac{x^m}{m^q} \int_0^1 t^{m-1} S_n(t)\,dt. \tag{8.6}$$

Using (8.5) and the series (2.5) and (2.9), we obtain

$$\phi(x) = \sum_{j=0}^{n-2} (-1)^j s_{n-j} \sum_{m=1}^{\infty} \frac{x^m}{m^{q+j+1}} - (-1)^n \sum_{m=1}^{\infty} \left(\sum_{k=1}^{m} \frac{1}{k} \right) \frac{x^m}{m^{q+n}} \tag{8.7}$$

$$= \sum_{j=0}^{n-2} (-1)^j s_{n-j} S_{q+j+1}(x) - (-1)^n \big(S_{n+q+1}(x) + S_{n+q-1,2}(x) \big).$$

Some other integrals containing $S_n(x)$ can be found in [12] and [15].

9. Values of $S_{n,p}(z)$ for special arguments z.

9.1. The integrals $s_{n,p}=S_{n,p}(1)$. As has been shown in [13], the homogeneous polynomial in $\zeta(q)$ for the integrals

$$s_{n,p}=\frac{(-1)^{n+p-1}}{(n-1)!p!}\int_0^1 \log^{n-1}t\log^p(1-t)\frac{dt}{t} \tag{9.1}$$

can be written in the form

$$s_{n,p}=(-1)^{n+p-1}\sum_{\nu=0}^{n-1} b_{n-\nu-1}\sum_{\rho=0}^{p}\binom{\nu+\rho+1}{\rho}b_{p-\rho}a_{\nu+\rho+1}$$

where [8, No. 8.321]

$$a_k=-\frac{1}{k}\sum_{m=1}^{k}(-1)^m\zeta(m)a_{k-m},\qquad b_k=\frac{1}{k}\sum_{m=1}^{k}(-1)^m\zeta(m)b_{k-m},$$

are the coefficients in the power series for $1/\Gamma(1+x)$ and $\Gamma(1+x)$ respectively, and where $a_0=b_0=1$ and $\zeta(1)\equiv\gamma$ by definition. In order to derive this expression, the well-known relation [8, No. 9.122 1]

$${}_2F_1(\alpha,\beta;\beta+1;1)=\frac{\Gamma(1-\alpha)\Gamma(1+\beta)}{\Gamma(1-\alpha+\beta)} \tag{9.2}$$

is essential.

9.2. The integrals $\sigma_{n,p}=(-1)^pS_{n,p}(-1)$. The problem of evaluating

$$\sigma_{n,p}=\frac{(-1)^{n-1}}{(n-1)!p!}\int_0^1 \log^{n-1}t\log^p(1+t)\frac{dt}{t} \tag{9.3}$$

seems to be far more difficult. This difficulty is related to the apparent nonexistence of a relation for ${}_2F_1(\alpha,\beta;\beta+1;-1)$ similar to (9.2). Nielsen claimed that he had succeeded in representing $\sigma_{n,p}$ by $\zeta(q)$ if $n+p$ is odd, but, as we shall see, his proof is erroneous. Those relations between $\sigma_{n,p}$ and $s_{n,p}$ which can be deduced from relations given in the previous sections are of such a special nature that it is not easy to see how they could be used, for example, to derive an expression for $\sigma_{n,p}$ in terms of $\zeta(k)$. On the contrary, it seems likely that, in the general case, no such relations exist. (Barlow [4] states, without giving a reference, that "$S_{n,p}(-1)$, $S_{n,p}(\frac{1}{2})$ and $S_{n,p}(1)$ are easily evaluated analytically from equation [(2.1)]".)

Nielsen did, however, find a closed formula for the case $p=2$, n odd, which seems to be the most general result known for $\sigma_{n,p}$. He proved the following

THEOREM 2. *Let $\zeta(k)$ be the Riemann zeta function for integer argumnent. Then for $n\geqq 0$*

$$\begin{aligned}\sigma_{2n+1,2}&=\frac{1}{2(2n)!}\int_0^1 \log^{2n}t\log^2(1+t)\frac{dt}{t}\\ &=\frac{1}{2}\left[1-(2n+1)(1-2^{-2n-2})\right]\zeta(2n+3)\\ &\quad+\sum_{j=1}^{n}(1-2^{1-2j})\zeta(2j)\zeta(2n-2j+3).\end{aligned} \tag{9.4}$$

The sum is zero if $n=0$.

Proof. Nielsen's proof, which is full of misprints, starts from the inversion formula (5.12) for $z=1$ and $p=2$. After applying Lemma (4.1a), he obtains an expression for $C_{2n-1,2}$. He then takes the real part and compares his result with the power series expansion of both sides of the relation

$$\pi \csc \pi x = \pi \cot \pi x \cos \pi x + \pi \sin \pi x.$$

Theorem 1 allows us to shorten this proof. From (5.9), we have

$$\sigma_{2n-1,2} = \frac{1}{2}\left[C_{2n-1,2} - (2n-1)\sigma_{2n+1}\right].$$

Hence, for $n=1$, using (7.1) and (2.22),

$$\sigma_{1,2} = \frac{1}{2}(C_{1,2} - \sigma_3) = \frac{1}{2}(s_3 - \sigma_3) = \frac{1}{8}\zeta(3).$$

For $n \geqq 2$, we obtain, using (7.2),

$$\sigma_{2n-1,2} = \frac{1}{2}\left\{2\sum_{j=0}^{n-1}(-1)^j \frac{\pi^{2j}}{(2j)!} s_{2n-2j-1,2} - (2n-1)\sum_{j=0}^{n-1}(-1)^j \frac{\pi^{2j}}{(2j)!} s_{2n-2j+1} - (2n-1)\sigma_{2n+1}\right\}. \tag{9.5}$$

With the help of [13], [20],

$$s_{m,2} = \frac{1}{2}\left[(m+1)s_{m+2} - \sum_{k=2}^{m} s_k s_{m-k+2}\right], \qquad (m \geqq 2) \tag{9.6}$$

$$s_{1,2} = s_3,$$

(9.5) can be written as

$$\sigma_{2n-1,2} = \frac{1}{2}\left\{-\sum_{j=0}^{n-1}(-1)^j \frac{\pi^{2j}}{2j(2j-2)!} s_{2n-2j+1} - \sum_{j=0}^{n-2}(-1)^j \frac{\pi^{2j}}{(2j)!} \sum_{k=2}^{2n-2j-1} s_k s_{2n-2j-k+1} + s_{2n+1} - (2n-1)\sigma_{2n+1}\right\}.$$

Noting the symmetry in the double sum, and reordering, gives

$$\sigma_{2n-1,2} = \frac{1}{2}\left\{\sum_{j=1}^{n-1}\left[(-1)^{j-1}\frac{\pi^{2j}}{2j(2j-2)!} - 2\sum_{k=0}^{j-1}(-1)^k \frac{\pi^{2k}}{(2k)!} s_{2j-2k}\right] \times s_{2n-2j+1} + s_{2n+1} - (2n-1)\sigma_{2n+1}\right\}.$$

Referring to (7.2), we see that the expression in square brackets equals $C_{2j-1,1} = 2\sigma_{2j}$, hence

$$\sigma_{2n-1,2} = \sum_{j=1}^{n-1} \sigma_{2j} s_{2n-2j+1} + \frac{1}{2}\left[s_{2n+1} - (2n-1)\sigma_{2n+1}\right], \tag{9.7}$$

and Theorem 2 follows on replacing n by $n+1$ and using (2.22).

Formula (9.4) shows that $\sigma_{2n+1,2}$ can be expressed by a homogeneous polynomial of degree $2n+3$ in $\zeta(q)$ $(2 \leqq q \leqq 2n+3)$, with rational coefficients. In contrast to the expressions for $s_{n,p}$ and $C_{n,p}$, however, all coefficients of terms which contain more than two factors vanish. A short table of $\sigma_{2n+1,2}$ is given in Table 2.

TABLE 2

Let

$$I_n = \sigma_{2n+1,2} = \frac{1}{2(2n)!}\int_0^1 \log^{2n} t \log^2(1+t)\frac{dt}{t}.$$

Then

$$I_0 = \frac{1}{8}\zeta(3)$$
$$I_1 = \frac{1}{2}\zeta(2)\zeta(3) - \frac{29}{32}\zeta(5)$$
$$I_2 = \frac{1}{2}\zeta(2)\zeta(5) + \frac{7}{8}\zeta(3)\zeta(4) - \frac{251}{128}\zeta(7)$$
$$I_3 = \frac{1}{2}\zeta(2)\zeta(7) + \frac{31}{32}\zeta(3)\zeta(6) + \frac{7}{8}\zeta(4)\zeta(5) - \frac{1529}{512}\zeta(9)$$
$$I_4 = \frac{1}{2}\zeta(2)\zeta(9) + \frac{127}{128}\zeta(3)\zeta(8) + \frac{7}{8}\zeta(4)\zeta(7) + \frac{31}{32}\zeta(5)\zeta(6) - \frac{8183}{2048}\zeta(11)$$
$$I_5 = \frac{1}{2}\zeta(2)\zeta(11) + \frac{511}{512}\zeta(3)\zeta(10) + \frac{7}{8}\zeta(4)\zeta(9) + \frac{127}{128}\zeta(5)\zeta(8) + \frac{31}{32}\zeta(6)\zeta(7) - \frac{40949}{8192}\zeta(13)$$

For a discussion of Nielsen's work on $\sigma_{n,p}$ for other values of n and p, it is perhaps useful to quote his monograph [20, pp. 199–202, translated from the German]. He starts with formula (2.22) for σ_n and continues that

> a similar result is not available for the more general numbers $\sigma_{n,p}\cdots$, if $p>1$. On the contrary, it seems that these numbers cannot, in general, be expressed in closed form by the numbers s_k. In any case, I have found such an expression only when $n+p$ is odd.

After having proved Theorem 2, he continues

> On the values $\sigma_{n,p}$, we still have to prove that, for $n \geqq 2$,
>
> (N1) $$\sigma_{p,n} = (-1)^{p-1}\binom{n+p-1}{n}\sigma_{1,n+p-1} + \sum_{q=0}^{p-2}(-1)^q\binom{n+q}{n}s_{n+q+1,p-q-1}.$$
>
> For this purpose, we set $x=1$ in (2) and obtain
>
> (N2) $$s_{n,p} = \sum_{q=0}^{p}\binom{n+p-q-1}{n-1}\sigma_{q+1,n+p-q-1},$$
>
> which easily gives formula (N1) by induction. Setting $n=2$ and $p=2n-1$ in (N1) yields
>
> (N3) $$\sigma_{2n-1,2} = \binom{2n+1}{2}\sigma_{1,2n} + \sum_{q=0}^{2n-3}(-1)^q\binom{q+2}{q}s_{q+3,2n-q-2}.$$
>
> Therefore, because of [(9.4)] and (N1), we can deduce the following theorem:
>
> For $n+p$ odd, the value $\sigma_{n,p}$ is a homogeneous polynomial of degree $n+p$ in $s_2, s_3, \cdots, s_{n+p}$ with rational coefficients, provided s_r is of degree r.
>
> In order to find the numbers $\sigma_{n,p}$ also for $n+p$ even, one needs only to express the numbers $\sigma_{1,2n-1}$ by s_r, in which I did not succeed.

A numerical check for small n and p shows, however, that (N2) cannot be correct as it stands. In addition, formula (N2), which represents, for given $n+p$, a system of $n+p-1$ linear equations in the $n+p-1$ unknowns $\sigma_{\nu,\rho}$, is not symmetric with respect

[2] There is an obvious misprint in the equation number referenced.

to n and p, as it should be according to (2.21). Further, for $n=1$ it contains a quantity $\sigma_{n+p,0}$ which is undefined. However, it is easy to see the origin of (N2) in spite of the misprint in the equation number referenced. It lies in Nielsen's incorrect equation corresponding to (6.6), which gives, for $a=0$, $b=1$, $z=1$, the relation (N2) as a special case.

A correct relation between $s_{n,p}$ and $\sigma_{\nu,\rho}$ is given by the following

THEOREM 3. *The numbers* $s_{n,p}=S_{n,p}(1)$ *and* $\sigma_{\nu,\rho}=(-1)^{\rho}S_{\nu,\rho}(-1)$ *are linearly related through*

$$s_{n,p}=\sum_{j=1}^{n}\binom{n+p-j-1}{p-1}\sigma_{j,n+p-j}+\sum_{j=1}^{p}\binom{n+p-j-1}{n-1}\sigma_{j,n+p-j}. \tag{9.8}$$

Proof. From (5.1), we obtain for $z=\frac{1}{2}$

$$L_{n,p}\left(\frac{1}{2}\right)+L_{p,n}\left(\frac{1}{2}\right)=s_{n,p}-\frac{\log^{n+p}2}{n!p!}$$

and from (6.2) for $z=1$, using (2.21),

$$L_{n,p}\left(\frac{1}{2}\right)+L_{p,n}\left(\frac{1}{2}\right)=2s_{n,p}-\left[\binom{n+p-1}{n-1}+\binom{n+p-1}{p-1}\right]\frac{\log^{n+p}2}{(n+p)!}$$
$$-\sum_{j=1}^{n}\binom{n+p-j-1}{p-1}\sigma_{j,n+p-j}-\sum_{j=1}^{p}\binom{n+p-j-1}{n-1}\sigma_{j,n+p-j}.$$

Theorem 3 follows by comparison.

Note that (9.8) is symmetric in n and p. This relation for $s_{n,p}$ was given in [12] without proof. Note also that setting $z=1$ in formula (6.7), derived from the correct equation (6.6) merely yields an equation equivalent to equation (9.8), as can be seen from elementary properties of binomial coefficients.

It seems puzzling that Nielsen should have published such an obviously incorrect equation as (N2), when the correct relation follows easily from his relations for $L_{n,p}(z)$.[3] However, by introducing (N1) into (N2) and performing some manipulations with binomial coefficients, it can be shown that (N1) is indeed the solution of the linear system (N2) for $\sigma_{\nu,\rho}$. This suggests that the more special formula (N3) for $\sigma_{1,2n}$ is also incorrect, a fact which is easily verified by a numerical check using small values of n, after correcting an obvious misprint in (N3).

By comparing (N2) and (9.8), we see that Nielsen was misled by the (wrong) triangular structure of (N2), which enabled him to solve his system in closed form. On the other hand, the correct system (9.8) does not allow a solution $\sigma_{n,p}$ in terms of $s_{\nu,\rho}$ (and therefore of $\zeta(q)$). For $n+p\geqq 4$, the system (9.8) consists of $[(n+p)/2]$ equations with either $n+p-2$ ($n+p$ even) or $n+p-3$ ($n+p$ odd) unknowns, so that for $n+p\geqq 6$ there are more unknowns than equations. The cases $n+p=4$ and $n+p=5$, where one has as many equations as unknowns, namely two, result in a vanishing determinant, yielding the single equations

$$2\sigma_{1,3}+\sigma_{2,2}=\frac{\pi^4}{720}, \qquad 2\sigma_{1,4}+\sigma_{2,3}=\frac{31}{32}\zeta(5)-\frac{1}{2}\zeta(2)\zeta(3). \tag{9.9}$$

The cases $n+p=2$ and $n+p=3$ are trivial.

[3] One could imagine that Nielsen, or somebody else at the time might have recognized this error later. However, a search through the volumes of *Nova Acta Leopoldina* and *Oversigt Danske Vidensk. Selsk. Forh.* up to 1918 shows no published erratum or remark relating to Nielsen's paper. As far as the author knows, no collected papers of Nielsen exist (cf. the list of known collected papers in Struik [22, p. 196]).

When examining other relations of §§5 and 6, one realizes that there seems to be an inherent difficulty in getting relations which could be used to compute $\sigma_{n,p}$ in terms of $s_{\nu,\rho}$. This is true in particular for the (singular) system (5.9) of linear equations for the unknowns $\sigma_{\nu,\rho}$. Therefore, it seems likely that, not only are Nielsen's relations (N1) and (N3) incorrect, but that no such relation exists. An unsuccessful computer search through relations of the form

$$\left.\begin{matrix}|\sigma_{1,4}|\\ |\sigma_{2,3}|\end{matrix}\right\} = \frac{p_1}{p_2}\zeta(2)\zeta(3)+\frac{p_3}{p_4}\zeta(5),$$

with integers p_i in the range $-50 \leqq p_1 \leqq 50$, $1 \leqq p_2 \leqq 50$, $0 \leqq p_3 \leqq 50$, $1 \leqq p_4 \leqq 50$ supports this conclusion.

9.3. The integrals $a_{n,p} = S_{n,p}(\frac{1}{2})$. The problem of evaluating

$$a_{n,p} = \frac{(-1)^{n+p-1}}{(n-1)!p!}\int_0^1 \log^{n-1} t \log^p\left(1-\frac{1}{2}t\right)\frac{dt}{t} \tag{9.10}$$

apparently faces obstacles similar to those found for the evaluation of $\sigma_{n,p}$; in particular, because no closed expression seems to be known for ${}_2F_1(\alpha,\beta;\beta+1;\frac{1}{2})$. A result which can easily be obtained from formulae treated in the previous sections is the following

THEOREM 4. *For $n \geqq 1$ and $p \geqq 1$, the numbers $a_{n,p}$ can be expressed in terms of $\sigma_{\nu,\rho}$ and $\log 2$, through the relation*

$$a_{n,p} = \sum_{j=0}^{n-1}(-1)^j\frac{\log^j 2}{j!}\sum_{k=1}^{n-j}\binom{n+p-j-k-1}{p-1}\sigma_{k,n+p-j-k} + (-1)^n\frac{\log^{n+p}2}{(n+p)!}. \tag{9.11}$$

Proof. From (6.2) with $z=1$ and (2.16) with $z=\frac{1}{2}$ it follows that

$$\sum_{i=0}^{n-1}\frac{\log^j 2}{j!}a_{n-j,p} = s_{n,p} - \frac{1}{(n+p)!}\binom{n+p-1}{n-1}\log^{n+p}2 - \sum_{k=1}^{\infty}\binom{n+p-k-1}{n-1}\sigma_{k,n+p-k}.$$

Applying Lemma (4.1a) and using (5.11) yields

$$a_{n,p} = \sum_{j=0}^{n-1}\frac{(-1)^j\log^j 2}{j!}\left\{s_{n-j,p} - \sum_{k=1}^{p}\binom{n+p-j-k-1}{n-j-1}\sigma_{k,n+p-j-k}\right\} + (-1)^n\frac{\log^{n+p}2}{(n+p)!}. \tag{9.12}$$

Replacing $s_{n-j,p}$ by (9.8) gives Theorem 4.

As a special case, we obtain for $p=1$,

$$a_n = \sum_{j=0}^{n-2}\frac{(-1)^j\log^j 2}{j!}\sum_{k=1}^{n-j-1}\sigma_{k,n-j-k} - (-1)^n\frac{\log^n 2}{n!}, \qquad (n \geqq 2). \tag{9.13}$$

In particular, for $n=2$,

$$a_2=-\int_0^1 \log\left(1-\frac{1}{2}t\right)\frac{dt}{t}=\sigma_{1,1}-\frac{1}{2}\log^2 2=\frac{\pi^2}{12}-\frac{1}{2}\log^2 2, \tag{9.14}$$

which is a well-known formula due to Euler (attributed to Landen by Lewin [17, p. 6]); and for $n=3$,

$$a_3=\int_0^1 \log t\log\left(1-\frac{1}{2}t\right)\frac{dt}{t}=\frac{7}{8}\zeta(3)-\frac{\pi^2}{12}\log 2+\frac{1}{6}\log^3 2. \tag{9.15}$$

For $n=4$, we have, using (9.9),

$$a_4=\frac{\pi^4}{90}+\frac{\pi^2}{24}\log^2 2-\frac{7}{8}\zeta(3)\log 2-\frac{1}{24}\log^4 2-\sigma_{1,3}. \tag{9.16}$$

The appearance of the quantity $\sigma_{1,3}$, for which no closed formula seem to be known, makes it impossible to compute

$$a_4=-\frac{1}{2}\int_0^1 \log^2 t\log\left(1-\frac{1}{2}t\right)\frac{dt}{t}$$

from (9.13).

From the power series (2.5) we have

$$a_n=\sum_{j=1}^{\infty}\frac{1}{2^j j^n}. \tag{9.17}$$

Therefore

$$a_2=\sum_{j=1}^{\infty}\frac{1}{2^j j^2}=\frac{\pi^2}{12}-\frac{1}{2}\log^2 2 \tag{9.18}$$

$$a_3=\sum_{j=1}^{\infty}\frac{1}{2^j j^3}=\frac{7}{8}\zeta(3)-\frac{\pi^3}{12}\log 2+\frac{1}{6}\log^3 2 \tag{9.19}$$

and, trivially, $a_0=1$, $a_1=\log 2$. Formula (9.19) had been found by Legendre.

It seems puzzling that (9.17) should exist in closed form only for $n=0,1,2,3$. Knowing a_4 for example would allow us to obtain, from (9.16) and (9.9), expressions for $\sigma_{1,3}$ and $\sigma_{2,2}$. This in turn, together with a knowledge of a_5, would make possible a computation of $\sigma_{1,4}$ and $\sigma_{2,3}$ from (9.12), (9.4) and (9.9). Levine et al. [15] and Gastmans and Troost [7] therefore consider a_4 and a_5 as "new" constants. This makes it possible to express more kinds of logarithmic integrals in "closed" form.

Another special case is $n=1$. Here we find from (9.11) that

$$a_{1,p}=\frac{(-1)^p}{p!}\int_0^1 \log^p\left(1-\frac{1}{2}t\right)\frac{dt}{t}=\sigma_{1,p}-\frac{\log^{p+1}2}{(p+1)!}. \tag{9.20}$$

Note that Nielsen's remark [20, p. 203] that $a_{1,2p}$ can be expressed in terms of $\log 2$ and $\zeta(q)$ $(2 \leqq q \leqq 2p+1)$, is probably incorrect, being a consequence of the incorrect formula (N3).

From (2.7), (9.4) and (9.20), one finds

$$a_{1,2} = \sum_{j=0}^{\infty} \frac{(-1)^j S_{j+2}^{(2)}}{(j+2)!(j+2)} \frac{1}{2^{j+2}}$$

$$= \sum_{j=2}^{\infty} \frac{1}{2^j j^2} \left(1 + \frac{1}{2} + \cdots + \frac{1}{j-1}\right) = \frac{1}{8}\zeta(3) - \frac{1}{6}\log^3 2$$

or

$$\int_0^1 \log^2\left(1 - \frac{1}{2}t\right) \frac{dt}{t} = \frac{1}{4}\zeta(3) - \frac{1}{3}\log^3 2. \tag{9.21}$$

Nielsen also gave two relations between $a_{1,n}$ and $a_{p,1}$. From (5.1) and (2.16), we obtain for $z = \frac{1}{2}$,

$$a_{1,n} = s_{n+1} - \sum_{j=0}^{n-1} \frac{\log^j 2}{j!} a_{n-j+1} - \frac{1}{n!}\log^{n+1} 2. \tag{9.22}$$

With the help of Lemma (4.1a) and a well-known relation for the binomial coefficients, (9.22) can be inverted to give

$$a_n = \sum_{j=0}^{n-2} \frac{(-1)^j \log^j 2}{j!} (s_{n-j} - a_{1,n-j-1}) - (-1)^n \frac{\log^n 2}{(n-1)!}. \tag{9.23}$$

Acknowledgment. I would like to thank Dr. B. Schorr for a critical reading of the manuscript and for helpful discussions.

REFERENCES

[1] M. Abramowitz and I. A. Stegun, editors, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, 5th printing with corrections, Nat. Bur. Standards Appl. Math. Series 55, US Government Printing Office, Washington, DC, 1966.

[2] R. Barbieri, J. A. Mignaco and E. Remiddi, *Electron form factors up to the fourth order*, I *and* II, Nuovo Cimento A, 11 (1972), pp. 824–864.

[3] R. Barbieri, M. Caffo, E. Remiddi, S. Turini and J. Oury, *The anomalous magnetic moment of the electron in* QED: *Some more sixth order contributions in the dispersive approach*, Nucl. Phys. B, 144 (1978), pp. 329–348.

[4] R. H. Barlow, *Convergent continued fraction approximants to generalized polylogarithms*, BIT, 14 (1974), pp. 112–116.

[5] L. Comtet, *Advanced Combinatorics*, Reichel, Dordrecht and Boston, 1974.

[6] J. A. Fox and A. C. Hearn, *Analytic computation of some integrals in fourth order quantum electrodynamics*, J. Comp. Phys., 14 (1974), pp. 301–317.

[7] R. Gastmans and W. Troost, *On the evaluation of polylogarithmic integrals*, Simon Stevin (Ghent), 55 (1981), pp. 205–219.

[8] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, Academic Press, New York, 1980.

[9] W. Gröbner and N. Hofreiter, *Integraltafel, zweiter Teil, Bestimmte Integrale*, Springer, Wien, 1973.

[10] A. C. Hearn, REDUCE *User's Manual, version* 3.0, Rand Publ. CP78, Santa Monica, CA, 1983.

[11] D. JACOBS AND F. LAMBERT, *On the numerical calculation of polylogarithms*, BIT, 12 (1972), pp. 581–585.
[12] K. S. KÖLBIG, J. A. MIGNACO AND E. REMIDDI, *On Nielsen's generalized polylogarithms and their numerical calculation*, BIT, 10 (1970), pp. 38–74.
[13] K. S. KÖLBIG, *Closed expressions for the integral* $\int_0^1 t^{-1}\log^{n-1} t \log^p(1-t)\,dt$, Math. Comp., 39 (1982), pp. 647–654.
[14] _____, *On the integral* $\int_0^{\pi/2} \log^n \cos x \log^p \sin x\,dx$, Math. Comp., 40 (1983), pp. 565–570.
[15] M. J. LEVINE, E. REMIDDI AND R. ROSKIES, *Analytic contributions to the g factor of the electron in sixth order*, Phys. Rev. D, 20 (1979), pp. 2068–2077.
[16] L. LEWIN, *Dilogarithms and Associated Functions*, Macdonald, London, 1958.
[17] _____, *Polylogarithms and Associated Functions*, North-Holland, New York, 1981.
[18] D. MAISON AND A. PETERMANN, *Subtracted generalized polylogarithms and the* SINAC *program*, Comput. Phys. Commun., 7 (1974), pp. 121–134.
[19] J. A. MIGNACO AND E. REMIDDI, *Fourth order vacuum polarization in the 6th order electron magnetic moment*, Nuovo Cimento A, 60 (1969), pp. 519–528.
[20] N. NIELSEN, *Der Eulersche Dilogarithmus und seine Verallgemeinerungen*, Nova Acta Leopoldina, 90 (1909), pp. 123–211.
[21] J. RIORDAN, *Combinatorial Identities*, John Wiley, New York, 1966.
[22] D. J. STRUIK, *Abriss der Geschichte der Mathematik*, VEB Deutscher Verlag der Wissenschaften, Berlin, 1976.

SIAM J. MATH. ANAL.
Vol. 17, No. 5, September 1986

014

p-ARY SEQUENCY AND ORDERINGS OF THE CHRESTENSON FUNCTIONS*

ZHANG GONGLI†

Abstract. A physical interpretation of the concept of the p-ary sequency is presented. A definition of p-ary sequency is given according to the interpretation. Expressions of p-ary natural (generalized Hadamard), p-ary sequency, and p-adic ordered Chrestenson functions are established in a systematic way, and they are related to the p-adic code and the p-ary Gray code. Mutual mappings of different ordered Chrestenson functions are discussed.

Key words. Chrestenson function, p-ary sequency, ordering, Gray code, Walsh function

1. Introduction. Much interest in multiple-valued devices, logic, and digital circuits has been evidenced in the past decade. The Chrestenson functions [1], which take multiple complex values, are compatible with p-ary numerical systems, and they have been applied to the design and optimization of multiple-valued logic networks [2], [3], [4]. In order to introduce the Chrestenson functions well to engineering, where p-ary digital techniques are employed, the present author showed that the sums of the real and imaginary parts of the Chrestenson functions also form a set of complete orthogonal functions [5]. We call them the RMV Walsh functions. Because the RMV Walsh functions take real multiple values, they are readily represented by physical means, e.g., voltage levels, as is necessary for system implementation. S-H Chang and T. Joseph [6] introduced the generalized concept of symmetry and sequence for the Chrestenson functions, and they presented orderings of the functions according to the symmetry and sequency. However, further development is necessary.

2. The p-ary sequency. For the Walsh functions the sequency, which is generalized frequency, is defined as "one-half the average number of zero crossings per unit of time" [7]. However, the definition is not applicable to the Chrestenson functions. In [6] sequency is generalized as the measure of oriented phase shift, but the physical interpretation of the concept is not given. The phase number or sequency number has not so obvious a meaning as the number of zero crossings.

It is known that for a given natural number m the Chrestenson functions [1] $\mathrm{CH}_w^p(t)$ are defined as

$$\mathrm{CH}_w^p(t)=\exp\left(2\pi i p^{-1}\mathbf{w}^T\mathbf{t}\right) \tag{1}$$

where

$$w=0,1,\cdots,p^{m-1},\quad w=\sum_{s=0}^{m-1} w_s p^{m-1-s},\quad \mathbf{w}=[w_0 w_1\cdots w_{m-1}]^T;$$

$$t\in[0,1),\quad t=\sum_{s=0}^{\infty} t_s p^{-s-1},\quad \mathbf{t}=[t_0 t_1\cdots t_{m-1}]^T,\quad i^2=-1;$$

$$J_p=\{0,1,\cdots,p-1\},\quad t_s, w_s\in J_p,\quad p \text{ is a fixed integer},\quad p>1.$$

*Received by the editors October 22, 1984. This work was supported in part by the Science Fund of the Chinese Academy of Science.

†Department of Information Engineering, Northwest Telecommunication Engineering Institute, Xi'an, China.

We would interpret variable t as time measured in seconds. To interpret the concept of sequency for Chrestenson functions, we begin by considering the physical interpretation of frequency for the exponential and trigonometric functions. Speaking more generally, we view that these functions are intimately related to the circular motions of particles, with which Newtonian (classical) mechanics deals. The exponential functions $\exp(2\pi iwt)$, $w=0,1,2,\cdots$, mathematically describe circular motion of a unit vector around the origin in the complex plane, as shown in Fig. 1. The unit vector projected onto the real axis yields $\cos(2\pi wt)$, while the projection onto the imaginary axis yields $\sin(2\pi wt)$. Whenever one uses the term frequency and angular shift of $\exp(2\pi iwt)$, $\sin(2\pi wt)$ and $\cos(2\pi wt)$, one refers implicitly to the circular motion described by these functions. The usual physical interpretation of frequency is the "number of cycles per unit time" [7]. It is just the definition of frequency, and its unit is "cps" or Hz. Note that the motion of the unit vector described by the exponential functions is continuous and differentiable, and the motion can be understood with the aid of Newtonian mechanics, which deals with forces, masses and motions of particles. The time t (function variable) is interpreted as the topology of continuum just as shown in the expressions for exponential functions. In other words, t has the same topology as the real numbers. From the point of view of mathematics the exponential functions represent a mapping of the real number t on the number axis in the interval $[0,1)$ onto the unit circle in the complex plane as shown in Fig. 1. The time t in the expressions for exponential functions is a continuous (analogue) variable.

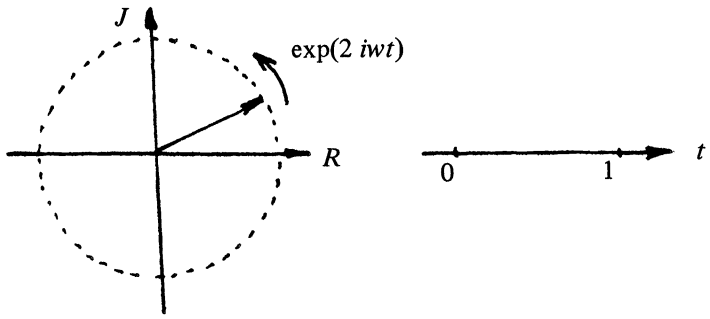


FIG. 1. *The motion of the unit vector described by the exponential functions.*

Similarly the Chrestenson functions may be said to describe mathematically the circular motion of a unit vector around the origin in the complex plane; but the circular motion described by the Chrestenson functions cannot be understood with the aid of Newtonian mechanics, if we view the motion as that of a particle. The unit vector transitionally moves discontinuously around the origin while the variable t increases from 0 to 1. The angular shift θ (in radians) of the position occupied by the unit vector is only an integer times the basis $\theta=2\pi/p$, i.e., $\theta=0,\ 2\pi/p,\ 2\times 2\pi/p,\cdots$. For example, if $p=3$ the three positions which the unit vector can occupy are shown in Fig. 2.

It is clear that the positions occupied by the unit vector are quantized. The unit vector moves instantaneously from one position to another without spending any time. The motion of the unit vector can be interpreted with the aid of quantum mechanics, which shows us that the transition from one position, form, or state to another is possible. The Chrestenson functions represent another mapping of the real number t on the number axis in the inverval $[0,1)$ onto the unit circle, but in the expression for

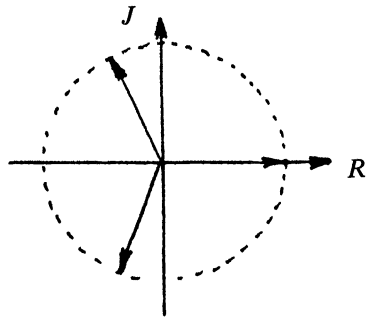


FIG. 2. *The three positions of the unit vector.*

Chrestenson functions the continuous time variable t, which has been quantized in the p-ary numerical system, becomes a digital variable. The discontinuous and transitional motion of the unit vector is related to the idea that the time t is regarded as having a p-adic topology for the Chrestenson functions.

The "cycle" for the Chrestenson functions thus has an obvious meaning according to the above interpretation. We interpret the sequency of a Chrestenson function as the "average number of cycles made by the rotating unit vector, the motion of which is described by the Chrestenson function, per unit of time". It is also the definition of p-ary sequency. The concept of p-ary sequency, which has a broader meaning, is applicable to the trigonometric, exponential, Walsh, and RMV Walsh functions if the function describing the motion of the unit vector is a corresponding function. The unit of p-ary sequency is the same as the one of frequency, i.e., "cps" or Hz. It is easy to see that p-ary sequency is a generalized concept of frequency and sequency.

We now show how the notion of p-ary sequency applies to the Walsh functions ($p=2$). Because $p=2$, the unit vector occupies only two positions, which are $\theta_1=2n\pi$ and $\theta_2=(2n+1)\pi$, $n=0,1,2,\cdots$. With θ_1 the value of the Walsh function equals 1, and with θ_2 its value is -1. If a Walsh function has one sign change (from 1 to -1 or the reverse) at t_1, $t_1\in[0,1)$, the corresponding angular shift of the unit vector equals π at t_1. Suppose N is the average number of sign changes (zero crossings) of the Walsh function per unit of time: according to the definition of p-ary sequency the sequency of the Walsh function would be

$$S=N\pi/2\pi=N/2. \tag{2}$$

It is just "one-half the average number of zero crossings per unit of time". Therefore, with $p=2$ the p-ary sequency and the sequency for the Walsh functions are identical.

It will be proved later that if $w=(w=0,1,\cdots,p^m-1)$ is the index of a p-ary sequency ordered Chrestenson function, the p-ary sequency s of the Chrestenson function is

$$S=[(p-1)w/p]^* \tag{3}$$

where $[a]^*$ denotes the least integer $\geqq a$.

3. Orderings of the Chrestenson functions. The question about the ordering of Walsh functions has been investigated by some authors [8], [9]. The merits and disadvantages of different orderings are discussed by Yuen [10]. There are three orderings for the Chrestenson functions, which correspond to the dyadic (Paley), natural

(Hadamard), and sequency (Walsh) ordering of Walsh functions. They are p-adic, p-ary natural (generalized Hadamard), and p-ary sequency.

Suppose w_H, w_G, and w_P express the indexes of the generalized Hadamard, p-ary sequency, and p-adic ordered Chrestenson functions respectively, and their p-adic code, p-ary Gray code [11], and bit-reversed p-adic code are represented as $(w_0 w_1 \cdots w_{m-1})$, $(g_0 g_1 \cdots g_{m-1})$, and $(h_0 h_1 \cdots h_{m-1})$, respectively. Then

$$h_s = w_{m-1-s}, \tag{4}$$

$$g_s = w_s \ominus w_{s-1}, \tag{5}$$

$$w_s = \sum_{k=0}^{s} g_k, \tag{6}$$

$$w_A = \sum_{s=0}^{m-1} w_s p^{m-1-s} \tag{7}$$

where $\ominus$ denotes subtraction modulo p and $\sum$ denotes summation modulo p.

$$A \in \{H, G, P\} \quad \text{and} \quad w_A = 0, 1, \cdots p^m - 1.$$

3.1. Generalized Hadamard ordering. The generalized Hadamard ordering is the one which was originally employed by H. E. Chrestenson [1]. The generalized Hadamard ordered Chrestenson functions $\mathrm{CH}^p_{w_H}(t)$ are expressible by using the bit-reversed p-adic code $(h_0 h_1 \cdots h_{m-1})$ of w_H as

$$\mathrm{CH}^p_{w_H}(t) = \exp\left((2\pi i/p) \sum_{s=0}^{m-1} h_{m-1-s} t_s \right). \tag{8}$$

Substituting (4) in (8), we find

$$\mathrm{CH}^p_{w_H}(t) = \exp\left((2\pi i/p) \sum_{s=0}^{m-1} w_s t_s \right). \tag{9}$$

If $[\mathrm{CH}_m]$ denotes the matrix whose w_Hth row is $[\mathrm{CH}^p_{w_H}(0)\ \mathrm{CH}^p_{w_H}(p^{-m}) \cdots \mathrm{CH}^p_{w_H}((p^m - 1)p^{-m})]$ for $w_H = 0, 1, \cdots, p^m - 1$, it is easily shown that [6]

$$[\mathrm{CH}_m] = [\mathrm{H}_m], \tag{10}$$

where $[\mathrm{H}_m]$ is the generalized Hadamard matrix of order m

$$[\mathrm{H}_1] = \begin{bmatrix} r^0 & r^0 & \cdots & r^0 \\ r^0 & r^1 & \cdots & r^{p-1} \\ \cdots & & & \\ r^0 & r^{p-1} & \cdots & r^1 \end{bmatrix}, \tag{11}$$

$$[\mathrm{H}_2] = [\mathrm{H}_1] \otimes [\mathrm{H}_1], \tag{12}$$

$$[\mathrm{H}_m] = [\mathrm{H}_{m-1}] \otimes [\mathrm{H}_1] \tag{13}$$

where $\otimes$ denotes the Kronecker product and $r = \exp(2\pi i/p)$.

3.2. *p*-Ary sequency ordering. The sequency ordered Chrestenson functions $\mathrm{CH}^p_{w_G}(t)$ are defined by using the p-ary Gray code $(g_0 g_1 \cdots g_{m-1})$ of w_G as

$$\mathrm{CH}^p_{w_G}(t) = \exp\left((2\pi i/p) \sum_{s=0}^{m-1} g_{m-1-s} t_s \right). \tag{14}$$

Substituting (5) in (14), we get

$$\text{(15)}\qquad \mathrm{CH}^{p}_{w_G}(t)=\exp\left((2\pi i/p)\sum_{s=0}^{m-1}(w_{m-1-s}\ominus w_{m-2-s})t_s\right).$$

Now we prove that the Chrestenson functions defined by (14) are ordered according to the p-ary sequency. As already mentioned, the p-ary sequency S of a Chrestenson function equals the angular shift (measured in radians) of the unit vector, the motion of which is described by the Chrestenson function over $[0,1)$, divided by 2π. For $\mathrm{CH}^{p}_{w_G}(t)$, the angular shift θ over $[0,1)$ is determined by

$$\text{(16)}\qquad \theta=\sum_{q=0}^{p^m-2}(2\pi/p)\left(\sum_{s=0}^{m-1}g_{m-1-s}t_{q+1,s}\ominus\sum_{s=0}^{m-1}g_{m-1-s}t_{q,s}\right)+\phi$$

where $q=\sum_{s=0}^{m-1}t_{q,s}p^{m-1-s}$, $t_{q,s}\in J_p$ and ϕ is the angular shift at $t=0$. We have

$$\text{(17)}\qquad \theta=(2\pi/p)\sum_{q=0}^{p^m-2}\left[\sum_{s=0}^{m-1}g_{m-1-s}(t_{q+1,s}\ominus t_{q,s})\right]+\phi.$$

According to the construction of the p-adic code of t, as shown in Table 1 for $p=3$, $m=3$, it is not difficult to see that

$$\text{(18)}\qquad \begin{aligned}\theta=(2\pi/p)\big[&(p^m-p^{m-1})g_0+(p^{m-1}-p^{m-2})(g_0\oplus g_1)\\&+\cdots+(p-1)(g_0\oplus g_1\oplus 2/\oplus g_{m-1})\big]+\phi,\end{aligned}$$

where $\oplus$ denotes addition modulo p. Substituting (6) in (18), we have

$$\text{(19)}\qquad \theta=(2\pi/p)(p-1)(p^{m-1}w_0+p^{m-2}w_1+\cdots+w_{m-1})+\phi$$

$$\text{(20)}\qquad =(2\pi/p)(p-1)\sum_{s=0}^{m-1}w_s p^{m-1-s}+\phi.$$

By (7),

$$\text{(21)}\qquad \theta=2\pi(p-1)w_G/p+\phi.$$

Therefore,

$$\text{(22)}\qquad S=\theta/2\pi=(p-1)w_G/p+\phi/2\pi.$$

TABLE 1

The values of $(t_{q+1}\ominus st_{q,s})$ *for* $p=3$, $m=3$.

t,q	t_0	t_1	t_2	$(t_{q+1}\ominus t_{q,s})$			t,q	t_0	t_1	t_2	$(t_{q+1}\ominus t_{q,s})$		
				$s=0$	1	2					$s=0$	1	2
0	0	0	0	0	0	1	14	1	1	2	0	1	1
1	0	0	1	0	0	1	15	1	2	0	0	0	1
2	0	0	2	0	1	1	16	1	2	1	0	0	1
3	0	1	0	0	0	1	17	1	2	2	1	1	1
4	0	1	1	0	0	1	18	2	0	0	0	0	1
5	0	1	2	0	1	1	19	2	0	1	0	0	1
6	0	2	0	0	0	1	20	2	0	2	0	1	1
7	0	2	1	0	0	1	21	2	1	0	0	0	1
8	0	2	2	1	1	1	22	2	1	1	0	0	1
9	1	0	0	0	0	1	23	2	1	2	0	1	1
10	1	0	1	0	0	1	24	2	2	0	0	0	1
11	1	0	2	0	1	1	25	2	2	1	0	0	1
12	1	1	0	0	0	1	26	2	2	2			
13	1	1	1	0	0	1							

The angular shift ϕ at $t=0$ would be determined, while the Chrestenson function is extended with period 1, i.e.,

$$\mathrm{CH}^{p}_{w_G}(t)=\mathrm{CH}^{p}_{w_G}(t+1). \tag{23}$$

It is easy to see that the ϕ is the angle shift from θ_1 (corresponding to $\mathrm{CH}^{p}_{w_G}(1-p^{-m})$) to θ_2 (corresponding to $\mathrm{CH}^{p}_{w_G}(0)$). So

$$S=[(p-1)w_G/p]^*. \tag{24}$$

The formula (24) shows that the larger the index w_G, the larger the corresponding p-ary sequency, and the $\mathrm{CH}^{p}_{w_G}(t)$ $(w_G=0,1,\cdots,p^m-1)$ are ordered according to the p-ary sequency. The p-ary sequencies of the Chrestenson functions for $p=3$, $m=3$, are shown in Table 2.

TABLE 2

The p-ary sequencies of the Chrestenson functions for $p=3$, $m=3$.

w_G	p-ary Gray code	$(\theta-\phi)$ (π)	S	w_G	p-ary Gray code	$(\theta-\phi)$ (π)	S
0	000	0	0	14	101	28	10
1	001	2	1	15	111	30	10
2	002	4	2	16	112	32	11
3	012	6	2	17	110	34	12
4	010	8	3	18	210	36	12
5	011	10	4	19	211	38	13
6	021	12	4	20	212	40	14
7	022	14	5	21	222	42	14
8	020	16	6	22	220	44	15
9	120	18	6	23	221	46	16
10	121	20	7	24	201	48	16
11	122	22	8	25	202	50	17
12	102	24	8	26	200	52	18
13	100	26	9				

3.3. *p*-Adic ordering. Using the p-adic code $(w_0w_1\cdots w_{m-1})$ of index w_p, we define the p-adic ordered Chrestenson functions $\mathrm{CH}^{p}_{w_p}(t)$ as

$$\mathrm{CH}^{p}_{w_p}(t)=\exp\left((2\pi i/p)\sum_{s=0}^{m-1} w_{m-1-s}t_s\right). \tag{25}$$

S-H Chang and T. Joseph [6] pointed out that the p-adic code $(w_0w_1\cdots w_{m-1})$ is the p-ary symmetry index for the $\mathrm{CH}^{p}_{w_p}(t)$, and that the vector

$$\left[\mathrm{CH}^{p}_{w_p}(0)\ \mathrm{CH}^{p}_{w_p}(p^{-m})\cdots\mathrm{CH}^{p}_{w_p}((p^m-1)p^{-m})\right]$$

can be written from left to right by using the p-ary symmetry index without referring to any other member of the set of Chrestenson functions. In [6] the p-adic ordered Chrestenson functions are called ordered according to the p-ary symmetry. We note that for the otherwise ordered Chrestenson functions the p-ary symmetry holds true, but the symmetry index is indifferent. The symmetry index of the generalized Hadamard ordered Chrestenson functions is the bit-reversed p-adic code, and the p-ary Gray code corresponds to the p-ary symmetry index of p-ary sequency ordered Chrestenson functions.

4. Mutual mappings of differently ordered Chrestenson functions. The exponential functions $\exp(2\pi iwt)$ are a character group of the topologic group of the real numbers. Since we may consider real numbers to be points on the real line which has only one dimension, the exponential functions have one ordering. On the other hand, the Chrestenson functions are a character group of the p-adic group, and the p-tuple vectors which belong to the p-adic group form a space which has p dimensions. Therefore, the Chrestenson functions can be ordered in many different ways.

Now we give a representation of the Chrestenson functions, ordering them according to the parameter matrix Q_A

$$\mathrm{CH}^{p}_{w_A}(t)=\exp\big((2\pi i/p)\mathbf{w}_A^T Q_A \mathbf{t}\big), \tag{26}$$

where A_A is an invertible symmetric matrix,

$$Q_A=[q_{ij}], \qquad i,j=0,\cdots,m-1, \quad q_{ij}\in J_P;$$
$$w_A=\sum_{s=0}^{m-1} w_s p^{m-1-s}, \quad \mathbf{w}_A=[w_0 w_1 \cdots w_{m-1}]^T, \quad \mathbf{t}=[t_0 t_1 \cdots t_{m-1}]^T.$$

All multiplication and addition operations in the calculation of $\mathbf{w}_A^T Q_A \mathbf{t}$ are performed modulo p. The ordering of the Chrestenson functions is determined by the matrix Q_A; the three matrices corresponding to the p-ary sequency, p-adic, and generalized Hadamard orderings are shown in Table 3.

TABLE 3

The matrices corresponding to the three orderings.

$$Q_H=\begin{bmatrix} 1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix} \qquad Q_G=\begin{bmatrix} & & & p-1 & 1 \\ & & \iddots & 1 & \\ & \iddots & \iddots & & \\ p-1 & \iddots & & & \\ 1 & & & & \end{bmatrix} \qquad Q_P=\begin{bmatrix} & & & & 1 \\ & & & \iddots & \\ & & 1 & & \\ & \iddots & & & \\ 1 & & & & \end{bmatrix}$$

Generalized Hadamard $\qquad$ p-ary sequency $\qquad$ p-adic

It is easy to see that the conversion from one ordering to another order is the same as the conversion from one code to the other code. Suppose $\mathbf{w}_H$, $\mathbf{w}_G$, and $\mathbf{w}_P$ are vectors formed by the p-adic code $(w_0 w_1 \cdots w_{m-1})$ of the indices w_H, w_G, and w_P, respectively. For example, if the p-adic code of w_H is (211), $\mathbf{w}_H=[211]^T$. The mutual mappings of the three orderings are represented as

$$\mathbf{w}_B^T=\mathbf{w}_A^T Q_A Q_B^{-1}, \tag{27}$$

where A, $B\in\{H,G,P\}$.

For example, let $p=3$, $m=3$; then

$$Q_H=\begin{bmatrix} 1 & & \\ & 1 & \\ & & 1 \end{bmatrix}, \quad Q_G=\begin{bmatrix} & 2 & 1 \\ 2 & 1 & \\ 1 & & \end{bmatrix}, \quad Q_P=\begin{bmatrix} & & 1 \\ & 1 & \\ 1 & & \end{bmatrix}$$

and

$$Q_H^{-1}=Q_H, \quad Q_P^{-1}=Q_P, \quad Q_G^{-1}=\begin{bmatrix} & & 1 \\ & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}.$$

Suppose $w_H = 8$; then $\mathbf{w}_H = [022]^T$,

$$\mathbf{w}_G^T = [022] \times \begin{bmatrix} 1 & & \\ & 1 & \\ & & 1 \end{bmatrix} \times \begin{bmatrix} & & 1 \\ & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} = \begin{bmatrix} & & 1 \\ & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} = [211],$$

$$\mathbf{w}_P^T = [022] \times \begin{bmatrix} 1 & & \\ & 1 & \\ & & 1 \end{bmatrix} \times \begin{bmatrix} & & 1 \\ & 1 & \\ 1 & & \end{bmatrix} = [220].$$

Using (7), we have $w_G = 22$, $w_P = 24$.

The mutual mappings of other Chrestenson functions are analogous to the mutual mappings of the foregoing three ordered functions.

5. Discussion. The Chrestenson functions correspond to the discontinuous circular motion of the unit vector around the origin in the complex plane. The p-ary sequency of the Chrestenson functions is analogous to frequency. Among the orderings of the functions there are three important orderings, which are generalized Hadamard, p-ary sequency and p-adic. The merits and disadvantages of the three orderings are analogous to those of the Walsh functions [10]. It is convenient to use the p-adic ordered Chrestenson functions for mathematical discussions, while the generalized Hadamard ordered functions are applicable to fast transform. p-Ary sequency ordered Chrestenson functions, the indices of which have a transparent physical interpretation, would be applied to communication engineering, p-ary sequency filtering, and so on, just as sequency ordered Walsh functions [7].

Acknowledgments. The author is grateful to Prof. Fan Changxin of Northwest Telecommunication Engineering Institute, China and Prof. Hu Zhengming of Beijing Institute of Posts and Telecommunications for their helpful suggestions.

REFERENCES

[1] H. F. CHRESTENSON, *A class of generalized Walsh functions*, Pacific J. Math., 5 (1955), pp. 17–31.

[2] M. G. KARPOVSKY, *Finite Orthogonal Series in the Design of Digital Devices*, John Wiley, New York, 1976.

[3] C. MORI[illegible]GA, *Complex spectral logic*, Proc. Eighth International Multiple-Valued Logic Conference, Rosemont, IL, 1976.

[4] GONGLI ZHANG, *The parameter spectrum in spectral multiple-valued logic design*, Electronics Lett., 19 (1983), pp. 199–200.

[5] _____, *Two complete orthogonal sets of real multiple-valued functions*, Proc. 14th International Symposium on Multiple-Valued Logic, Winnipeg, Manitoba, Canada, 1984, pp. 12–18.

[6] S-H CHANG AND T. JOSEPH, *On ordering of generalized Walsh functions*, Proc. 1972 Symposium on Applications of Walsh functions, Naval Research Lab., Washington, DC, pp. 337–343.

[7] H. F. HARMUTH, *A generalized concept of frequency and some applications*, IEEE Trans. Inform. Theory, IT-14 (1968), pp. 375–382.

[8] K. W. HENDERSON, *Some notes on the Walsh functions*, IEEE Trans. Electron, Comput., EC-13 (1964), pp. 50–52.

[9] H. KREMER, *Representations and mutual relations of the different systems of Walsh functions*, Theory and Applications of Walsh Functions and Other Non-Sinusoidal Functions, The Hatfield Polytechnic, 1973.

[10] C. K. YUEN, *Remarks in the orderings of Walsh functions*, IEEE Trans. Comput., C-21 (1972), p. 1452.

[11] BHU DEV SHARMA AND RAVINDER KUMAR KNANNA, *On m-ary Gray codes*, Inform. Sci. (1978), pp. 31–43.

SIAM J. MATH. ANAL.
Vol. 17, No. 5, September 1986

015

ANOTHER CONJECTURED q-SELBERG INTEGRAL*

MIZAN RAHMAN†

Abstract. Using Askey and Wilson's q-beta type integral, a q-extension of Selberg's n-dimensional integral is given as a conjecture and is proved for $n=2$.

Key words. Selberg's integral, Askey–Wilson integral, Sears' summation formula, basic hypergeometric series, quadratic transformation-formula, Macdonald–Morris conjectures

AMS(MOS) subject classifications. Primary 33A15, 33A65

1. Introduction. In a little-known paper published in a Norwegian journal, A. Selberg [14] gave the following important result:

$$(1.1)\qquad \int_{-1}^{1}\cdots\int_{-1}^{1}\prod_{i=1}^{n}(1-x_i)^{\alpha}(1+x_i)^{\beta}\prod_{1\le i\le j\le n}|x_j-x_i|^{2\gamma}\,dx_1\cdots dx_n$$

$$=\prod_{j=1}^{n}2^{\alpha+\beta+1+(n-1)\gamma}\frac{\Gamma(\alpha+1+(j-1)\gamma)\Gamma(\beta+1+(j-1)\gamma)\Gamma(j\gamma+1)}{\Gamma(\alpha+\beta+2+(n+j-2)\gamma)\Gamma(\gamma+1)},$$

$$\operatorname{Re}\alpha>-1,\quad \operatorname{Re}\beta>-1,\quad \operatorname{Re}\gamma>-\left[\frac{1}{n},\frac{\operatorname{Re}(\alpha+1)}{n-1},\frac{\operatorname{Re}(\beta+1)}{n-1}\right],\quad n\ge 2.$$

In view of the particular extension we wish to propose, we have expressed the integrals on the interval $[-1,1]$ instead of $[0,1]$ and this explains an additional factor $2^{\alpha+\beta+1+(n-1)\gamma}$ on the right-hand side.

Once an important result is found, it is natural to look for extensions. Selberg's integral is a mutidimensional beta integral, so there may be extensions in each of the ways that Euler's beta integral has been extended. One way to extend the beta integral is as a sum, either like the hypergeometric distribution on an equally spaced set of points, or on a set of points whose ratio is constant. Included in the latter is Thomae's q-integral representation of the q-gamma function and Ramanujan's ${}_1\psi_1$ sum. See [15], [2]. Askey [3] discovered some extensions of Selberg's integral of this type, and was able to prove some of them in two dimensions. In a different direction, Macdonald [8] conjectured some constant term identities that are truncated versions of some of his earlier identities, Morris [9] made some further conjectures, and Macdonald reformulated Morris' conjectures so they made sense for the affine root systems [8]. For affine BC_n Macdonald's conjecture is equivalent to an integral with an absolutely continuous measure that extends Selberg's integral. For BC_1 Askey [4] proved this conjecture and the related integral was a new extension of Euler's integral. There is a

*Received by the editors July 10, 1984, and in revised form February 17, 1985.
†Department of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada, K1S 5B6.
This research was supported by the Natural Sciences and Engineering Research Council of Canada under grant A6197.

similar extension of Euler's integral with four free parameters in addition to q, which was found by Askey and Wilson [5]:

(1.2)

$$\int_{-1}^{1} \frac{h(x;1)h(x;-1)h(x;q^{1/2})h(x;-q^{1/2})}{h(x;a)h(x;b)h(x;c)h(x;d)} \frac{dx}{\sqrt{1-x^2}}$$

$$= \frac{2\pi(abcd;q)_\infty}{(q;q)_\infty(ab;q)_\infty(ac;q)_\infty(ad;q)_\infty(bc;q)_\infty(bd;q)_\infty(cd;q)_\infty}$$

$$= \kappa(a,b,c,d), \text{ say,}$$

where

$$(a;q)_\infty = \prod_{n=0}^{\infty}(1-aq^n), \qquad |q|<1, \tag{1.3}$$

and

$$h(x;a) = \prod_{n=0}^{\infty}(1-2axq^n+a^2q^{2n}) = (ae^{i\theta};q)_\infty(ae^{-i\theta};q)_\infty, \qquad x=\cos\theta, \tag{1.4}$$

the parameters being subject to the restriction

$$\max(|q|,|a|,|b|,|c|,|d|)<1. \tag{1.5}$$

It is natural to ask if this can be used to extend the Selberg integral. That (1.2) is an extension of Euler's integral can be seen by specializing the parameters in a number of different ways and then taking the limit $q\to 1$. Following Askey and Wilson [5], we take

$$0<q<1, \quad a=q^{\alpha/2+1/4}, \quad b=aq^{1/2}, \quad c=-q^{\beta/2+1/4}, \quad d=cq^{1/2}, \quad \alpha,\beta>-1. \tag{1.6}$$

Then, using the notation

$$(a;q)_\alpha = \frac{(a;q)_\infty}{(aq^\alpha;q)_\infty}, \tag{1.7}$$

(1.2) can be written as

(1.8)

$$\int_{-1}^{1} \left|(e^{i\theta};q)_{\alpha/2+1/4}(e^{i\theta}q^{1/2};q)_{\alpha/2+1/4}(-e^{i\theta};q)_{\beta/2+1/4}(-e^{i\theta}q^{1/2};q)_{\beta/2+1/4}\right|^2 \frac{dx}{\sqrt{1-x^2}}$$

$$= \frac{2\pi(-q^{1/2};q)_{(\alpha+\beta)/2}(-q;q)^2_{(\alpha+\beta)/2}(-q^{1/2};q)_{(\alpha+\beta+2)/2}}{\Gamma_q^2(1/2)} \frac{\Gamma_q(\alpha+1)\Gamma_q(\beta+1)}{\Gamma_q(\alpha+\beta+2)}.$$

In (1.8) we have used the identity

$$(q^{1/2};q)_\infty(-q^{1/2};q)_\infty(-q;q)_\infty = 1 \tag{1.9}$$

and the q-gamma function

$$\Gamma_q(x)=\frac{(q;q)_\infty}{(q^x;q)_\infty}(1-q)^{1-x},\quad \lim_{q\to 1}\Gamma_q(x)=\Gamma(x),\quad 0<q<1. \tag{1.10}$$

Observing that $\lim_{q\to 1}(z;q)_a=\lim_{q\to 1}\Sigma_{n=0}^{\infty}(q^{-a};q)_n(zq^a)^n/(q;q)_n=(1-z)^a$, and so

$$\lim_{q\to 1}\left|(e^{i\theta};q)_a\right|^2=2^a(1-\cos\theta)^a,\qquad \lim_{q\to 1}(-q^a;q)_b=2^b, \tag{1.11}$$

it is easily seen that (1.8) goes to the familiar beta integral in the limit $q\to 1$.

One can also see by similar arguments that

$$\lim_{q\to 1}\left|(q^{1/2}e^{i\theta+i\phi};q)_\gamma(q^{1/2}e^{i\theta-i\phi};q)_\gamma\right|^2=2^{2\gamma}|\cos\theta-\cos\phi|^{2\gamma}, \tag{1.12}$$

where γ is a real number. This provides the important clue to a possible extension of the discriminant $\Pi_{1\le i<j\le n}|x_j-x_i|^{2\gamma}$ in (1.1). Note that the limit on the right-hand side of (1.12) would be the same if we replaced $q^{1/2}$ by an arbitrary power of q. It turns out, however, that the integrals do not reduce to a compact formula without the $q^{1/2}$. It seems reasonable, therefore, to start with the following q-extension of the left-hand side of (1.1)

$$\int_{-1}^{1}\cdots\int_{-1}^{1}\prod_{j=1}^{n}w\left(x_j;a,aq^{1/2},b,bq^{1/2}\right)\cdot\prod_{1\le k<l\le n}\left|(q^{1/2}e^{i\theta_k+i\theta_l};q)_\gamma(q^{1/2}e^{i\theta_k-i\theta_l};q)_\gamma\right|^2 dx_1\cdots dx_n, \tag{1.13}$$

where

$$w\left(x_j;a,aq^{1/2},b,bq^{1/2}\right)=\frac{h(x_j;1)h(x_j;-1)h\left(x_j;q^{1/2}\right)h\left(x_j;-q^{1/2}\right)}{h(x_j;a)h\left(x_j;aq^{1/2}\right)h(x_j;b)h\left(x_j;bq^{1/2}\right)}\left(1-x_j^2\right)^{-1/2},\qquad x_j=\cos\theta_j. \tag{1.14}$$

At the moment we are unable to handle the integral for a γ that is not a nonnegative integer. So we shall assume that $\gamma=N$, $N=0,1,2,\cdots$. We will show that

$$\begin{aligned}\int_{-1}^{1}\int_{-1}^{1}w(x;a,aq^{1/2},b,bq^{1/2})w(y;a,aq^{1/2},b,bq^{1/2})&\cdot\left|(q^{1/2}e^{i\theta+i\phi};q)_N(q^{1/2}e^{i\theta-i\phi};q)_N\right|^2dx\,dy\\&=\prod_{j=1}^{2}\frac{2\pi}{(q;q)_\infty^2(1-ab)(abq^{1/2+(j-1)N};q)_\infty^2(abq^{1+(j-1)N};q)_\infty^2}\\&\cdot\prod_{j=1}^{2}\frac{(a^2b^2q^{jN+1};q)_\infty(q^{N+1};q)_\infty(q;q)_\infty}{(a^2q^{1/2+(j-1)N};q)_\infty(b^2q^{1/2+(j-1)N};q)_\infty(q^{jN+1};q)_\infty},\end{aligned} \tag{1.15}$$

where $x=\cos\theta$, $y=\cos\phi$.

We shall carry out the computations in two steps. In §2 we shall do integration over y and complete the second integration over x in §3. In §4 we shall state the n-dimensional result as a conjecture and examine the situation when we replace the

parameters as in (1.6) but also replace $e^{i\theta_j}$ by q^{ix_j} in the integrand and then take the limit $q\to 1$.

2. Evaluation of a single integral. We shall start by computing a somewhat more general integral

(2.1)

$$\int_{-1}^{1} w(y;a,b,c,d)(zq^{1/2}e^{i\phi};q)_r(zq^{1/2}e^{-i\phi};q)_r(q^{1/2}e^{i\phi}/z;q)_s(q^{1/2}e^{-i\phi}/z;q)_s\,dy$$
$$=S_{r,s}(z;a,b,c,d), \text{ say,}$$

where z is arbitrary and r,s are nonnegative integers. The formulas that we shall immediately need are the q-integral representations of Sears' summation formula [12], [1], [7]

(2.2)

$$\int_a^b \frac{(qu/a;q)_\infty(qu/b;q)_\infty(cu;q)_\infty}{(fu;q)_\infty(gu;q)_\infty(hu;q)_\infty}\,d_qu$$
$$=\frac{b(1-q)(q;q)_\infty(bq/a;q)_\infty(a/b;q)_\infty(c/f;q)_\infty(c/g;q)_\infty(c/h;q)_\infty}{(af;q)_\infty(ag;q)_\infty(ah;q)_\infty(bf;q)_\infty(bg;q)_\infty(bh;q)_\infty}$$

where $c=abfgh$, and of Bailey's formula [6, 8.5(3)] for the sum of two balanced and nonterminating balanced ${}_4\phi_3$'s in terms of a very well-poised ${}_8\phi_7$:

(2.3)

$$\int_a^b \frac{(qu/a;q)_\infty(qu/b;q)_\infty(cu;q)_\infty(du;q)_\infty}{(eu;q)_\infty(fu;q)_\infty(gu;q)_\infty(hu;q)_\infty}\,d_qu$$
$$=\frac{b(1-q)(q;q)_\infty(bq/a;q)_\infty(a/b;q)_\infty}{(ae;q)_\infty(af;q)_\infty(ag;q)_\infty}$$
$$\cdot\frac{(cd/eh;q)_\infty(cd/fh;q)_\infty(cd/gh;q)_\infty(bc;q)_\infty(bd;q)_\infty}{(be;q)_\infty(bf;q)_\infty(bg;q)_\infty(bh;q)_\infty(bcd/h;q)_\infty}$$
$$\cdot{}_8\phi_7\left[\begin{matrix}bcd/hq, q(bcd/hq)^{1/2}, -q(bcd/hq)^{1/2}, be, bf, bg, c/h, d/h\\ (bcd/hq)^{1/2}, -(bcd/hq)^{1/2}, cd/eh, cd/fh, cd/gh, bd, bc\end{matrix};q,\frac{cd}{befg}\right],$$

where $cd=abefgh$; see also [7, (3.25)].

The q-integrals in (2.2) and (2.3) are defined by

(2.4)
$$\int_0^a f(u)\,d_qu=a(1-q)\sum_{n=0}^{\infty}f(aq^n)q^n,$$
$$\int_a^b f(u)\,d_qu=\int_0^b f(u)\,d_qu-\int_0^a f(u)\,d_qu,$$

and the basic hypergeometric series ${}_{r+1}\phi_r$ by

(2.5)
$${}_{r+1}\phi_r\left[\begin{matrix}a_1,a_2,\cdots,a_{r+1}\\ b_1,\cdots,b_r\end{matrix};q,z\right]=\sum_{n=0}^{\infty}\frac{(a_1;q)_n(a_2;q)_n\cdots(a_{r+1};q)_n}{(q;q)_n(b_1;q)_n\cdots(b_r;q)_n}z^n,$$

provided $|z|<1$ if the series does not terminate.

A third formula that we shall find useful is the integral representation of an ${}_8\phi_7$ found by Nassrallah and Rahman [10]

$$(2.6)\quad \int_{-1}^{1} (w;\lambda,\mu,\nu,\rho)\frac{h(x;\tau)}{h(x;\sigma)}\,dx$$

$$=\frac{2\pi(\lambda\mu\nu\rho;q)_\infty(\lambda\mu\nu\sigma;q)_\infty(\lambda\tau;q)_\infty}{(q;q)_\infty(\lambda\mu;q)_\infty(\lambda\nu;q)_\infty(\lambda\rho;q)_\infty(\lambda\sigma;q)_\infty(\mu\nu;q)_\infty(\mu\rho;q)_\infty}$$

$$\cdot\frac{(\mu\tau;q)_\infty(\nu\tau;q)_\infty}{(\mu\sigma;q)_\infty(\nu\rho;q)_\infty(\nu\sigma;q)_\infty(\lambda\mu\nu\tau;q)_\infty}$$

$$\cdot{}_8\phi_7\left[\begin{matrix}\lambda\mu\nu\tau q^{-1}, q(\lambda\mu\nu\tau q^{-1})^{1/2}, -q(\lambda\mu\nu\tau q^{-1})^{1/2}, \lambda\mu, \lambda\nu, \mu\nu, \tau\rho^{-1}, \tau\sigma^{-1} \\ (\lambda\mu\nu\tau q^{-1})^{1/2}, -(\lambda\mu\nu\tau q^{-1})^{1/2}, \nu\tau, \mu\tau, \lambda\tau, \lambda\mu\nu\rho, \lambda\mu\nu\sigma\end{matrix};q,\rho\sigma\right],$$

see also [11].

Using (2.2), we find that

$$(2.7)\quad \frac{h(y;q^{1/2}/z)}{h(y;d)h(y;fq^{1/2}/z)}$$

$$=\frac{(f^{-1};q)_\infty(q^{1/2}/zd;q)_\infty}{fq^{1/2}z^{-1}(1-q)(q;q)_\infty(dz/fq^{1/2};q)_\infty(fq^{3/2}/dz;q)_\infty(dfq^{1/2}/z;q)_\infty}$$

$$\cdot\int_d^{fq^{1/2}/z}\frac{(qu/d;q)_\infty(uzq^{1/2}/f;q)_\infty(uq^{1/2}/z;q)_\infty}{(u/df;q)_\infty h(y;u)}\,d_qu.$$

Hence

$$(2.8)\quad \int_{-1}^{1} w(y;a,b,c,d)(ze^{i\phi}q^{1/2};q)_r(ze^{-i\phi}q^{1/2};q)_r\frac{h(y;q^{1/2}/z)}{h(y;fq^{1/2}/z)}\,dy$$

$$=\frac{(f^{-1};q)_\infty(q^{1/2}/dz;q)_\infty}{fq^{1/2}z^{-1}(1-q)(q;q)_\infty(dz/fq^{1/2};q)_\infty(fq^{3/2}/dz;q)_\infty(dfq^{1/2}/z;q)_\infty}$$

$$\cdot\int_d^{fq^{1/2}/z}\frac{(qu/d;q)_\infty(uzq^{1/2}/f;q)_\infty(uq^{1/2}/z;q)_\infty}{(u/df;q)_\infty}\,d_qu$$

$$\cdot\int_{-1}^{1} w(y;a,b,c,u)\frac{h(y;zq^{1/2})}{h(y;zq^{r+1/2})}\,dy.$$

By (2.6) we can now express the y-integral on the right-hand side as a very well-poised ${}_8\phi_7$ which terminates because of the assumption that r is a nonnegative integer, and consequently expressible as a balanced and terminating ${}_4\phi_3$ via Watson's formula [6, 8.5(2)]. Thus the y-integral on the right of (2.8) simplifies to

$$(2.9)\quad \frac{2\pi(azq^{1/2};q)_r(zq^{1/2}/a;q)_r(abcu;q)_\infty}{(q;q)_\infty(ab;q)_\infty(ac;q)_\infty(bc;q)_\infty(au;q)_\infty(bu;q)_\infty(cu;q)_\infty}$$

$$\cdot{}_4\phi_3\left[\begin{matrix}q^{-r}, ab, ac, au \\ azq^{1/2}, aq^{1/2-r}/z, abcu\end{matrix};q,q\right].$$

Substituting this into (2.8) the right-hand side becomes

(2.10)

$$\frac{2\pi(f^{-1};q)_\infty(q^{1/2}/dz;q)_\infty(azq^{1/2};q)_r(zq^{1/2}/a;q)_r}{fq^{1/2}z^{-1}(1-q)(q;q)_\infty^2(ab;q)_\infty(ac;q)_\infty(bc;q)_\infty(dz/fq^{1/2};q)_\infty(fq^{3/2}/dz;q)_\infty(dfq^{1/2}/z;q)_\infty}$$

$$\cdot\sum_{j=0}^{r}\frac{(q^{-r};q)_j(ab;q)_j(ac;q)_j}{(q;q)_j(azq^{1/2};q)_j(aq^{1/2-r}/z;q)_j}q^j$$

$$\cdot\int_d^{fq^{1/2}/z}\frac{(qu/d;q)_\infty(uzq^{1/2}/f;q)_\infty(uq^{1/2}/z;q)_\infty(abcuq^j;q)_\infty}{(bu;q)_\infty(cu;q)_\infty(u/df;q)_\infty(auq^j;q)_\infty}d_qu$$

$$=\frac{2\pi(bq^{1/2}/z;q)_\infty(cq^{1/2}/z;q)_\infty(qf/z^2;q)_\infty}{(q;q)_\infty(ab;q)_\infty(ac;q)_\infty(bc;q)_\infty(cd;q)_\infty}$$

$$\cdot\frac{(bcdfq^{1/2}/z;q)_\infty(azq^{1/2};q)_r(zq^{1/2}/a;q)_r}{(dfq^{1/2}/z;q)_\infty(bfq^{1/2}/z;q)_\infty(cfq^{1/2}/z;q)_\infty(bcfq/z^2;q)_\infty}$$

$$\cdot\sum_{j=0}^{r}\frac{(q^{-r};q)_j(ab;q)_j(ac;q)_j(abcfq^{1/2+j}/z;q)_\infty}{(q;q)_j(azq^{1/2};q)_j(aq^{1/2-r}/z;q)_j(afq^{1/2+j}/z;q)_\infty}q^j$$

$$\cdot{}_8\phi_7\left[\begin{matrix}bcf/z^2,q(bcf/z^2)^{1/2},-q(bcf/z^2)^{1/2},bfq^{1/2}/z,\\(bcf/z^2)^{1/2},-(bcf/z^2)^{1/2},cq^{1/2}/z,bq^{1/2}/z,\end{matrix}\right.$$

$$\left.\begin{matrix}cfq^{1/2}/z,q^{1/2}/dz,bc,q^{1/2-j}/az\\bcdfq^{1/2}/z,qf/z^2,abcfq^{1/2+j}\end{matrix};q,adq^j\right],$$

by (2.3). We now replace f by q^s, use [6, 8.5(3)] and simplify to get

(2.11)

$$S_{r,s}(z;a,b,c,d)$$

$$=\kappa(a,b,c,d)\frac{(azq^{1/2};q)_r(zq^{1/2}/a;q)_r(bq^{1/2}/z;q)_s(cq^{1/2}/z;q)_s(ad;q)_s}{(abcd;q)_s}$$

$$\cdot\sum_{j=0}^{r}\frac{(q^{-r};q)_j(ab;q)_j(ac;)_j(adq^s;q)_j}{(q;q)_j(azq^{1/2};q)_j(aq^{1/2-r}/z;q)_j(abcdq^s;q)_j}q^i$$

$$\cdot{}_4\phi_3\left[\begin{matrix}q^{-s},bc,q^{1/2}/dz,q^{1/2-j}/az\\bq^{1/2}/z,cq^{1/2}/z,q^{1-j-s}/ad\end{matrix};q,q\right].$$

The ${}_4\phi_3$ series on the right is balanced and terminating, so we may apply Sears' transformation formula [13]

(2.12)

$${}_4\phi_3\left[\begin{matrix}q^{-n},a,b,c\\d,e,f\end{matrix};q,q\right]$$

$$=\frac{(aq^{1-n}/e;q)_n(aq^{1-n}/f;q)_n}{(e;q)_n(f;q)_n}\left(\frac{bc}{d}\right)^n{}_4\phi_3\left[\begin{matrix}q^{-n},a,d/b,d/c\\d,aq^{1-n}/e,aq^{1-n}/f\end{matrix};q,q\right],$$

$def=abcq^{1-n}$, as often as necessary. Application of this formula twice in an order that should be obvious from the reduction, yields the following:

(2.13)

$$
{}_4\phi_3\left[\begin{matrix} q^{-s}, bc, q^{1/2}/dz, q^{1/2-j}/az \\ bq^{1/2}/z, cq^{1/2}/z, q^{1-j-s}/ad \end{matrix}; q, q\right]
$$

$$
= \frac{(aq^{1/2+j}/z;q)_s (q^{1/2}/az;q)_s (azq^{1/2};q)_j (abcdq^j;q)_s}{(bq^{1/2}/z;q)_s (cq^{1/2}/z;q)_s (azq^{1/2-s};q)_j (adq^j;q)_s} q^{-js}
$$

$$
\cdot {}_4\phi_3\left[\begin{matrix} q^{-s}, abq^j, acq^j, adq^j \\ abcdq^j, aq^{1/2+j}/z, azq^{1/2-s+j} \end{matrix}; q, q\right].
$$

Substituting this into (2.11) and simplifying the coefficients, we obtain

(2.14)

$$
S_{r,s}(z;a,b,c,d)
$$

$$
= \kappa(a,b,c,d)(azq^{1/2};q)_r (zq^{1/2}/a;q)_r (aq^{1/2}/z;q)_s (q^{1/2}/az;q)_s
$$

$$
\cdot \sum_{j=0}^{r} \sum_{k=0}^{s} \frac{(q^{-r};q)_j (q^{-s};q)_k (aq^{1/2+s}/z;q)_j (ab;q)_{j+k} (ac;q)_{j+k} (ad;q)_{j+k}}{(q;q)_j (q;q)_k (aq^{1/2-r}/z;q)_j (abcd;q)_{j+k} (aq^{1/2}/z;q)_{j+k} (azq^{1/2-s};q)_{j+k}}
$$

$$
\cdot q^{j+k-js}.
$$

We now transform the summation variables by setting $j+k=l$, $k=l-j$. The double sum in (2.14) becomes

(2.15)

$$
\sum_{l=0}^{r+s} \frac{(q^{-s};q)_l (ab;q)_l (ac;q)_l (ad;q)_l}{(q;q)_l (abcd;q)_l (aq^{1/2}/z;q)_l (azq^{1/2-s};q)_l} q^l {}_3\phi_2\left[\begin{matrix} q^{-l}, q^{-r}, aq^{1/2+s}/z \\ aq^{1/2-r}/z, q^{1+s-l} \end{matrix}; q, q\right]
$$

$$
= \sum_{l=0}^{r+s} \frac{(q^{-s};q)_l (ab;q)_l (ac;q)_l (ad;q)_l}{(q;q)_l (abcd;q)_l (aq^{1/2}/z;q)(azq^{1/2-s};q)_l} q^l \frac{(aq^{1/2}/z;q)_l (q^{-r-s};q)_l}{(aq^{1/2-r}/z;q)_l (q^{-s};q)_l}
$$

$$
= {}_4\phi_3\left[\begin{matrix} q^{-r-s}, ab, ac, ad \\ abcd, aq^{1/2-r}/z, azq^{1/2-s} \end{matrix}; q, q\right],
$$

by Jackson's formula [6, 8.4(1)] for the sum of a balanced and terminating ${}_3\phi_2$ series. Thus we have

(2.16)

$$
S_{r,s}(z;a,b,c,d)
$$

$$
= \kappa(a,b,c,d)(azq^{1/2};q)_r (zq^{1/2}/a;q)_r (aq^{1/2}/z;q)_s (q^{1/2}/az;q)_s
$$

$$
\cdot {}_4\phi_3\left[\begin{matrix} q^{-r-s}, ab, ac, ad \\ abcd, aq^{1/2-r}/z, azq^{1/2-s} \end{matrix}; q, q\right].
$$

Reversing the series on the right and simplifying, we find

(2.17)

$$S_{r,s}(z;a,b,c,d)=\kappa(a,b,c,d)\frac{(ab;q)_{r+s}(ac;q)_{r+s}(ad;q)_{r+s}}{(abcd;q)_{r+s}}a^{-r-s}z^{r-s}q^{(r+s)/2-rs}$$
$$\cdot{}_4\phi_3\left[\begin{matrix}q^{-r-s},q^{1-r-s}/abcd,zq^{1/2-s}/a,q^{1/2-r}/az\\ q^{1-r-s}/ab,q^{1-r-s}/ac,q^{1-r-s}/ad\end{matrix};q,q\right]$$
$$=\kappa(a,b,c,d)\frac{(ab;q)_{r+s}(bc;q)_{r+s}}{(abcd;q)_{r+s}}(bd;q)_{r+s}b^{-r-s}z^{r-s}q^{(r+s)/2-rs}$$
$$\cdot{}_{\phi 3}\left[\begin{matrix}q^{-r-s},q^{1-r-s}/abcd,zq^{1/2-s}/b,q^{1/2-r}/bz\\ q^{1-r-s}/ba,q^{1-r-s}/bc,q^{1-r-s}/bd\end{matrix};q,q\right],$$

the last line following from the previous one by (2.12). Replacing b,c,d by $aq^{1/2}$, b, $bq^{1/2}$, respectively, we finally obtain

(2.18)

$$S_{r,s}(z;a,aq^{1/2},b,bq^{1/2})$$
$$=\kappa(a,aq^{1/2},b,bq^{1/2})\frac{(a^2q^{1/2};q)_{r+s}(abq^{1/2};q)_{r+s}(abq;q)_{r+s}}{(a^2b^2q;q)_{r+s}}a^{-r-s}z^{r-s}q^{-rs}$$
$$\cdot{}_4\phi_3\left[\begin{matrix}q^{-r-s},q^{-r-s}/a^2b^2,zq^{-s}/a,q^{-r}/az\\ q^1z-r-s/a^2,q^{1/2-r-s}/ab,q^{-r-s}/ab\end{matrix};q,q\right].$$

The ${}_4\phi_3$ series on the right of (2.18) has a structure that enables us to apply a quadratic transformation formula due to Askey and Wilson [5, (4.22)]

$$(2.19)\qquad {}_4\phi_3\left[\begin{matrix}q^{-2n},a^2q^{2n},b^2q,c^2\\ -a,-aq,b^2c^2q^2\end{matrix};q^2,q^2\right]={}_4\phi_3\left[\begin{matrix}q^{-n},aq^n,b^2q,c^2\\ -a,bcq,-bcq\end{matrix};q,q\right].$$

Use of this in (2.18) followed by yet another application of (2.12) and some simplifications leads to

(2.20)

$$S_{r,s}(z;a,aq^{1/2},b,bq^{1/2})$$
$$=\kappa(a,aq^{1/2},b,bq^{1/2})\frac{(a^2q^{1/2};q)_{r+s}(abq^{1/2};q)_{r+s}(abq;q)_{r+s}}{(a^2b^2q;q)_{r+s}}$$
$$\cdot\frac{(bq^{1/4};q^{1/2})_{r+s}(-q^{1/2};q^{1/2})_{r+s}}{(-aq^{1/4};q^{1/2})_{r+s}(abq^{(r+s+1)/2};q^{1/2})_{r+s}}z^{r-s}q^{(r+s)^2/4-rs}$$
$$\cdot{}_4\phi_3\left[\begin{matrix}q^{-(r+s)/2},-q^{-(r+s)/2}/ab,q^{1/4+(s-r)/2}/z,zq^{1/4+(r-s)/2}\\ -q^{1/2},q^{1/4-(r+s)/2}/a,q^{1/4-(r+s)/2}/b\end{matrix};q^{1/2},q^{1/2}\right].$$

3. Proof of (1.15). It should be clear from (2.1) and (1.15) that we now need to evaluate the integral

$$\int_{-1}^{1} w(x;a,aq^{1/2},b,bq^{1/2})S_{N,N}(e^{i\theta};a,aq^{1/2},b,bq^{1/2})\,dx \equiv I_N, \quad \text{say.} \tag{3.1}$$

From (2.20) we have

(3.2)

$$S_{N,N}(e^{i\theta};a,aq^{1/2},b,bq^{1/2}) = \kappa(a,aq^{1/2},b,bq^{1/2}) \cdot \frac{(a^2q^{1/2},q)_{2N}(abq^{1/2};q)_{2N}(abq;q)_{2N}(bq^{1/4};q^{1/2})_{2N}(-q^{1/2};q^{1/2})_{2N}}{(a^2b^2q;q)_{2N}(-aq^{1/4};q^{1/2})_{2N}(abq^{N+1/2};q^{1/2})_{2N}} \cdot {}_4\phi_3\left[\begin{matrix} q^{-N}, -q^{-N}/ab, q^{1/4}e^{i\theta}, q^{1/4}e^{-i\theta} \\ q^{1/4-N}/a, q^{1/4-N}/b, -q^{1/2} \end{matrix}; q^{1/2}, q^{1/2}\right].$$

So we need to consider the integral

$$\int_{-1}^{1} w(x;a,aq^{1/2},b,bq^{1/2})(q^{1/4}e^{i\theta};q^{1/2})_m(q^{1/4}e^{-i\theta};q^{1/2})_m\,dx \equiv T_m, \quad \text{say,} \quad m=0,1,2,\cdots,N. \tag{3.3}$$

Using the identity

$$(a;q)_{2n} = (a;q^2)_n(aq;q^2)_n, \tag{3.4}$$

we find that when m is even

$$\left|(q^{1/4}e^{i\theta};q^{1/2})_m\right|^2 = \left|(q^{1/2}e^{i\theta};q)_{m/2}(q^{3/4}e^{i\theta};q)_{m/2}\right|^2, \tag{3.5}$$

and when m is odd

$$\left|(q^{1/4}e^{i\theta};q^{1/2})_m\right|^2 = \left|(q^{1/4}e^{i\theta};q)_{(m+1)/2}(q^{3/4}e^{i\theta};q)_{(m-1)/2}\right|^2. \tag{3.6}$$

Using (2.20), (3.4) and the identity

$$(a^2;q^2)_n = (a;q)_n(-a;q)_n, \tag{3.7}$$

we can easily show that

$$\begin{aligned} T_m &= S_{m/2,m/2}(q^{-1/4};a,aq^{1/2},b,bq^{1/2}), \quad m \text{ even} \\ &= S_{(m+1)/2,(m-1)/2}(q^{-1/4};a,aq^{1/2},b,bq^{1/2}), \quad m \text{ odd} \\ &= \kappa(a,aq^{1/2},b,bq^{1/2})(aq^{1/4};q^{1/2})_m(bq^{1/4};q^{1/2})_m(-q^{1/2};q^{1/2})_m \\ &\qquad (-abq^{1/2};q^{1/2})_m. \end{aligned} \tag{3.8}$$

From (3.1), (3.2), (3.3) and (3.8) we then have

$$I_N = \kappa^2(a,aq^{1/2},b,bq^{1/2})\frac{(a^2q^{1/2};q)_{2N}(abq^{1/2};q)_{2N}(abq;q)_{2N}}{(a^2b^2q;q)_{2N}} \cdot \frac{(bq^{1/4};q^{1/2})_{2N}(-q^{1/2};q^{1/2})_{2N}}{(-aq^{1/4};q^{1/2})_{2N}(abq^{N+1/2};q^{1/2})_{2N}} \cdot {}_4\phi_3\left[\begin{matrix} q^{-N}, -q^{-N}/ab, aq^{1/4}, bq^{1/4} \\ -abq^{1/2}, q^{1/4-N}/a, q^{1/4-N}/b \end{matrix}; q^{1/2}, q^{1/2}\right]. \tag{3.9}$$

This ${}_4\phi_3$ is summable. We have

(3.10)

$$
\begin{aligned}
&{}_4\phi_3\left[\begin{matrix} q^{-N}, -q^{-N}/ab, aq^{1/4}, bq^{1/4} \\ -abq^{1/2}, q^{1/4-N}/a, q^{1/4-N}/b \end{matrix}; q^{1/2}, q^{1/2}\right] \\
&\quad = \frac{(a^2q^{1/2};q^{1/2})_{2N}(abq^{1/2};q^{1/2})_{2N}}{(q^{1/4-N}/a;q^{1/2})_{2N}(q^{1/4-N}/b;q^{1/2})_{2N}}\left(\frac{q^{-N-1/4}}{a^2b}\right)^{2N} \\
&\qquad \cdot {}_4\phi_3\left[\begin{matrix} q^{-N}, a^2b^2q^{N+1/2}, aq^{1/4}, -aq^{1/4} \\ a^2q^{1/2}, abq^{1/2}, -abq^{1/2} \end{matrix}; q^{1/2}, q^{1/2}\right], \qquad \text{by (2.12)} \\
&\quad = \frac{(a^2q^{1/2};q^{1/2})_{2N}(abq^{1/2};q^{1/2})_{2N}}{(q^{1/4-N}/a;q^{1/2})_{2N}(q^{1/4-N}/b;q^{1/2})_{2N}}\left(\frac{q^{-N-1/4}}{a^2b}\right)^{2N} \\
&\qquad \cdot {}_4\phi_3\left[\begin{matrix} q^{-N}, a^2q^{1/2}, a^2q^{1/2}, a^2b^2q^{N+1/2} \\ a^2q^{1/2}, a^2q, a^2b^2q \end{matrix}; q, q\right] \qquad \text{by (2.19)} \\
&\quad = \frac{(a^2q^{1/2};q^{1/2})_{2N}(abq^{1/2};q^{1/2})_{2N}}{(q^{1/4-N}/a;q^{1/2})_{2N}(q^{1/4-N}/b;q^{1/2})_{2N}}\left(\frac{q^{-N-1/4}}{a^2b}\right)^{2N} \\
&\qquad \cdot \frac{(q^{1/2};q)_N(q^{1/2-N}/b^2;q)_N}{(a^2q;q)_N(q^{-N}/a^2b^2;q)_N} \\
&\quad = \frac{(a^2q^{1/2};q^{1/2})_{2N}(abq^{1/2};q^{1/2})_{2N}}{(aq^{1/4};q^{1/2})_{2N}(bq^{1/4};q^{1/2})_{2N}} \cdot \frac{(b^2q^{1/2};q)_N(q^{1/2};q)}{(a^2b^2q;q)_N(a^2q;q)_N}.
\end{aligned}
$$

Substituting this into (3.9), using (3.4), (3.7) and the definition of κ given in (1.2), we obtain (1.15).

4. The conjecture for the *n*-dimensional integral and the limiting cases. Our conjecture for the n-dimensional integral (1.13) is now clear from the form of (1.15):

$$
\begin{aligned}
(4.1)\quad &\int_{-1}^{1}\cdots\int_{-1}^{1}\prod_{j=1}^{n} w(x_j;a,aq^{1/2},b,bq^{1/2}) \\
&\qquad \cdot \prod_{1\le k<l\le n}\left|(q^{1/2}e^{i(\theta_k+\theta_l)};q)_N(q^{1/2}e^{i(\theta_k-\theta_l)};q)_N\right|^2 dx_1dx_2\cdots dx_n \\
&= \prod_{j=1}^{n}\frac{2\pi}{(q;q)_\infty^2(1-ab)(abq^{1/2+(j-1)N};q)_\infty^2(abq^{1+(j-1)N};q)_\infty^2} \\
&\qquad \cdot \prod_{j=1}^{n}\frac{(a^2b^2q^{1+(n+j-2)N};q)_\infty(q^{N+1};q)_\infty(q;q)_\infty}{(a^2q^{1/2+(j-1)N};q)_\infty(b^2q^{1/2+(j-1)N};q)_\infty(q^{jN+1};q)_\infty}.
\end{aligned}
$$

If we now take $0<q<1$, set $a=q^{\alpha/2+1/4}, b=-q^{\beta/2+1/4}$ and use (1.7), (1.9) and (1.10), we can also express it in the form

(4.2)

$$\int_{-1}^{1}\cdots\int_{-1}^{1}\prod_{j=1}^{n}\left|\left(e^{i\theta_j};q\right)_{\alpha/2+1/4}\left(q^{1/2}e^{i\theta_j};q\right)_{\alpha/2+1/4}\left(-e^{i\theta_j};q\right)_{\beta/2+1/4}\left(-q^{1/2}e^{i\theta_j};q\right)_{\beta/2+1/4}\right|^2\left(1-x_j^2\right)^{-1/2}$$

$$\cdot\prod_{1\le k<l\le n}\left|\left(q^{1/2}e^{i(\theta_k+\theta_l)};q\right)_N\left(q^{1/2}e^{i(\theta_k-\theta_l)};q\right)_N\right|^2 dx_1\,dx_2\cdots dx_n$$

$$=\prod_{j=1}^{n}\frac{2\pi\left(-q^{1/2};q\right)_{N(j-1)+(\alpha+\beta+2)/2}^2\left(-q;q\right)_{N(j-1)+(\alpha+\beta)/2}^2}{\Gamma_q^2(1/2)\left(1+q^{(\alpha+\beta+1)/2}\right)}$$

$$\cdot\prod_{j=1}^{n}\frac{\Gamma_q(\alpha+1+(j-1)N)\Gamma_q(\beta+1+(j-1)N)\Gamma_q(jN+1)}{\Gamma_q(\alpha+\beta+2+(n+j-2)N)\Gamma_q(N+1)}.$$

In this form it is clear that the limit of (4.2) is the Selberg formula (1.1). It is also clear from (4.2) that we have a conjecture for the constant term in the expansion of

$$\prod_{i=1}^{n}\left(x_i;q\right)_a\left(1/x_i;q\right)_a\left(-x_i;q\right)_b\left(-1/x_i;q\right)_b$$

$$\cdot\prod_{1\le i<j\le n}\left(qx_ix_j;q^2\right)_c\left(q/x_ix_j;q^2\right)_c\left(qx_i/x_j;q^2\right)_c\left(qx_j/x_i;q^2\right)_c.$$

However, there is no need for us to take b negative. We may, in fact, express the integral on the left of (4.1) in the form

$$\int_0^{\pi}\cdots\int_0^{\pi}\prod_{j=1}^{n}\left|\frac{\left(e^{2i\theta_j};q\right)}{\left(ae^{i\theta_j};q^{1/2}\right)_\infty\left(be^{i\theta_j};q^{1/2}\right)_\infty}\right|^2$$

$$\cdot\prod_{1\le j<k\le n}\left|\left(q^{1/2}e^{i(\theta_j+\theta_k)};q\right)_N\left(q^{1/2}e^{i(\theta_j-\theta_k)};q\right)_N\right|^2 d\theta_1\,d\theta_2\cdots d\theta_n,$$

so that if we now set

(4.3) $$a=q^{\alpha/2+1/4},\qquad b=q^{\beta/2+1/4},\qquad 0<q<1,$$

and transform the integral by setting $e^{i\theta_j}=q^{ix_j}$, that is, $\theta_j=x_j\log q$, then (4.1) gives

(4.4)

$$(\log q)^n\int_0^{\pi/\log q}\cdots\int_0^{\pi/\log q}\prod_{j=1}^{n}\left|\frac{\left(q^{2ix_j};q\right)_\infty}{\left(q^{\alpha/2+1/4+ix_j};q^{1/2}\right)_\infty\left(q^{\beta/2+1/4+ix_j};q^{1/2}\right)_\infty}\right|^2$$

$$\prod_{1\le j<k\le n}\left|\left(q^{1/2+ix_j+ix_k};q\right)_N\left(q^{1/2+ix_j-ix_k};q\right)_N\right|^2 dx_1\,dx_2\cdots dx_n$$

$$=\prod_{j=1}^{n}\frac{2\pi}{(q;q)_\infty^2\left(1-q^{(\alpha+\beta+1)/2}\right)\left(q^{(\alpha+\beta+2)/2+(j-1)N};q\right)_\infty^2\left(q^{(\alpha+\beta+3)/2+(j-1)N};q\right)_\infty^2}$$

$$\cdot\prod_{j=1}^{n}\frac{\left(q^{\alpha+\beta+2+(n+j-2)N};q\right)_\infty\left(q^{N+1};q\right)_\infty(q;q)_\infty}{\left(q^{\alpha+1+(j-1)N};q\right)_\infty\left(q^{\beta+1+(j-1)N};q\right)_\infty\left(q^{jN+1};q\right)_\infty}.$$

Using (1.10), this can also be written as

(4.5)

$$\left(\frac{\log q}{q-1}\right)^n (-1)^n \int_0^{\pi/\log q} \cdots \int_0^{\pi/\log q} \cdot \prod_{j=1}^{n} \left| \frac{\Gamma_q(\alpha/2+1/4+ix_j)\Gamma_q(\alpha/2+3/4+ix_j)\Gamma_q(\beta/2+1/4+ix_j)\Gamma_q(\beta/2+3/4+ix_j)}{\Gamma_q(2ix_j)} \right|^2$$

$$\cdot \prod_{1\le j<k\le n} \left| \frac{\Gamma_q(N+1/2+ix_j+ix_k)\Gamma_q(N+1/2+ix_j-ix_k)}{\Gamma_q(1/2+ix_j+ix_k)\Gamma_q(1/2+ix_j-ix_k)} \right|^2 dx_1 dx_2 \cdots dx_n$$

$$= \prod_{j=1}^{n} \frac{2\pi(1-q)\Gamma_q^2((\alpha+\beta+2)/2+(j-1)N)\Gamma_q^2((\alpha+\beta+3)/2+(j-1)N)}{(1-q^{(\alpha+\beta+1)/2})}$$

$$\cdot \prod_{j=1}^{n} \frac{\Gamma_q(\alpha+1+(j-1)N)\Gamma_q(\beta+1+(j-1)N)\Gamma q(jN+1)}{\Gamma_q(\alpha+\beta+2+(n+j-2)N)\Gamma_q(N+1)}.$$

If we now take the limit $q\to 1$ and use the duplication formula for the gamma function, we get the Mellin–Barnes type multiple integral

(4.6)

$$\int_0^\infty \cdots \int_0^\infty \prod_{j=1}^{n} \left| \frac{\Gamma(\alpha+1/2+2ix_j)\Gamma(\beta+1/2+2ix_j)}{\Gamma(2ix_j)} \right|^2$$

$$\cdot \prod_{1\le j<k\le n} \left|(1/2+ix_j+ix_k)_N (1/2+ix_j-ix_k)_N\right|^2 dx_1 dx_2 \cdots dx_n$$

$$= 2^{-2n(1+(n-1)N)} \prod_{j=1}^{n} \frac{\Gamma^2(\alpha+\beta+2+2(j-1)N)\Gamma(\alpha+1+(j-1)N)}{(\alpha+\beta+1)\Gamma(N+1)\Gamma(\alpha+\beta+2+(n+j-2)N)}$$

$$\frac{\Gamma(\beta+1+(j-1)N)\Gamma(jN+1)}{(\alpha+\beta+1)\Gamma(N+1)\Gamma(\alpha+\beta+2+(n+j-2)N)}$$

as a conjecture which is true if (4.1) is.

Acknowledgment. I am grateful to Professor R. Askey for making many important suggestions that I have freely used throughout the paper and for sending me a copy of Dyson's translation of Selberg's paper.

REFERENCES

[1] W. A. Al-Salem and A. Verma, *Some remarks on q-beta integral*, Proc. Amer. Math. Soc., 85 (1982), pp. 360–362.

[2] R. Askey, *Ramanujan's extensions of the gamma and beta functions*, Amer. Math. Monthly, 87 (1980), pp. 346–359.

[3] ———, *Some basic hypergeometric extensions of integrals of Selberg and Andrews*, this Journal, 11 (1980), pp. 938–951.

[4] ———, *A q-beta integral associated with BC_1*, this Journal, 13 (1982), pp. 1008–1010.

[5] R. ASKEY AND J. WILSON, *Some basic hypergeometric orthogonal polynomials that generalize Jacobi polynomials*, Mem. Amer. Math. Soc., 54, 319, 1985.

[6] W. N. BAILEY, *Generalized Hypergeometric Series*, Cambridge Univ. Press, Cambridge, 1935.

[7] G. GASPER AND MIZAN RAHMAN, *Positivity of the Poisson kernel for the continuous q-Jacobi polynomials and some quadratic transformation formulas for basic hypergeometric series*, to appear.

[8] I. G. MACDONALD, *Some conjectures for root systems*, this Journal, 13 (1982), pp. 988–1007.

[9] W. G. MORRIS, II, *Constant term identities for finite and affine root systems: conjectures and theorems*, Ph. D. thesis, Univ. Wisconsin, Madison, 1982.

[10] B. NASSRALLAH AND MIZAN RAHMAN, *Projection formulas, a reproducing kernel and a generating function for q-Wilson polynomials*, this Journal, 16 (1985), pp. 186–197.

[11] MIZAN RAHMAN, *An integral representation of a* ${}_{10}\phi_9$ *and continuous bi-orthogonal* ${}_{10}\phi_9$ *rational functions*, to appear.

[12] D. B. SEARS, *Transformations of basic hypergeometric functions of special type*, Proc. Lond. Math. Soc., 52 (1951), pp. 467–483.

[13] _____, *On the transformation theory of basic hypergeometric functions*, Proc. Lond. Math. Soc. (2), 53 (1951), pp. 158–180.

[14] A. SELBERG, *Bermerkninger om et Multipelt Integral*, Norsk Mat. Tidsskr., 26 (1944), pp. 71–78.

[15] J. THOMAE, *Beiträge zur Theorie der durch die Heinesche Reihe*: $1+((1-q^{\alpha})(1-q^{\beta})/(1-q)(1-q^{\gamma}))x + \cdots$ *daarstellbaren Functionen*, J. Reine Angew. Math., 70 (1869), pp. 258–281.

SIAM J. MATH. ANAL.
Vol. 17, No. 5, September 1986

016

q-WILSON FUNCTIONS OF THE SECOND KIND*

MIZAN RAHMAN†

Abstract. The q-Wilson function of the second kind is defined as a Hilbert transform of the q-Wilson polynomials and is shown to be expressible as a very well-poised ${}_8\phi_7$ series. This is used to find the leading terms in the asymptotic expansions of the q-Wilson polynomials and functions of the second kind on the interval $-1<x<1$.

Key words. q-Wilson polynomials, functions of the second kind, very-well-poised basic hypergeometric series, asymptotic properties

AMS(MOS) subject classifications. Primary 33A15, 33A65

1. Introduction. Let $\{p_n(x)\}_{n=0}^{\infty}$ be a set of polynomials orthogonal with respect to a positive measure $d\alpha(x)$ on the real line having support within a finite interval $[a,b]$. Then a corresponding function of the second kind is defined by

$$q_n(z)=\int_a^b \frac{p_n(t)}{z-t}\,d\alpha(t), \qquad z\notin[a,b]. \tag{1.1}$$

We are interested in the functions that correspond to the q-Wilson polynomials defined by

$$p_n\equiv p_n(x;a,b,c,d)={}_4\phi_3\left[\begin{matrix}q^{-n},abcdq^{n-1},ae^{i\theta},ae^{-i\theta}\\ ab,ac,ad\end{matrix};q,q\right], \tag{1.2}$$

where $-1\leqq x=\cos\theta\leqq 1$, and the basic hypergeometric series on the right is given by

$${}_{r+1}\phi_r\left[\begin{matrix}a_1,a_2,\cdots,a_{r+1}\\ b_1,\cdots,b_r\end{matrix};q,x\right]=\sum_{k=0}^{\infty}\frac{(a_1,a_2,\cdots,a_{r+1};q)_k}{(q,b_1,\cdots,b_r;q)_k}x^k, \tag{1.3}$$

with

$$(a_1,a_2,\cdots,a_j;q)_k=(a_1;q)_k(a_2;q)_k\cdots(a_j;q)_k, \tag{1.4}$$

$$(a;q)_k=\frac{(a;q)_\infty}{(aq^k;q)_\infty}, \qquad (a;q)_\infty=\prod_{n=0}^{\infty}(1-aq^n). \tag{1.5}$$

Askey and Wilson [1] showed that

$$\int_{-1}^{1} w(x)p_m(x)p_n(x)\,dx=h_n^{-1}\delta_{m,n}, \tag{1.6}$$

*Received by the editors May 17, 1984, and in revised form, February 1, 1985.
†Department of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada, K1S 5B6.
This research was supported by the Natural Sciences and Engineering Research Council of Canada under grant A6197.

where

$$w(x)=(1-x^2)^{-1/2}\frac{h(x;1)h(x;-1)h(x;\sqrt{q})h(x;-\sqrt{q})}{h(x;a)h(x;b)h(x;c)h(x;d)}, \tag{1.7}$$

$$h(x;a)=\prod_{n=0}^{\infty}(1-2axq^n+a^2q^{2n})=(ae^{i\theta},ae^{-i\theta};q)_\infty, \tag{1.8}$$

$$h_n=\kappa^{-1}(a,b,c,d)\frac{(abcdq^{-1};q)_n(1-abcdq^{2n-1})(ab,ac,ad;q)_n}{(q;q)_n(1-abcdq^{-1})(cd,bd,bc;q)_n}a^{-2n}, \tag{1.9}$$

with

$$\kappa(a,b,c,d)=\frac{2\pi(abcd;q)_\infty}{(q,ab,ac,ad,bc,bd,cd;q)_\infty}, \tag{1.10}$$

and

$$\max(|q|,|a|,|b|,|c|,|d|)<1. \tag{1.11}$$

In §2 we shall compute $q_n(z)$ by using (1.2) and $d\alpha(x)=w(x)dx$. Once the computation is done for $z\notin[-1,1]$ we shall be able to define an appropriate q-Wilson function of the second kind for $-1<x<1$ by making use of the well-known properties:

$$q_n(x)=\frac{1}{2}[q_n(x+i0)+q_n(x-i0)], \tag{1.12}$$

$$p_n(x)w(x)=-\frac{1}{2\pi i}[q_n(x+i0)-q_n(x-i0)]. \tag{1.13}$$

In §3 we shall discuss some asymptotic properties of $p_n(x)$ and $q_n(x)$ when $x\neq\pm1$.

2. q-Wilson function of the second kind. In (1.1) let us replace z by $\frac{1}{2}(z+z^{-1})=x$, say, and assume that $|z|<1$. If $t=\cos\phi$, then we may write

$$x-\cos\phi=\frac{1}{2z}(1-ze^{i\phi})(1-ze^{-i\phi})=\frac{1}{2z}\frac{h(t;z)}{h(t;qz)}. \tag{2.1}$$

Hence

$$\begin{aligned} q_n(x)&=2z\int_{-1}^{1}w(t;a,b,c,d)\frac{h(t;qz)}{h(t;z)}p_n(t;a,b,c,d)\,dt \\ &=2z\sum_{k=0}^{n}\frac{(q^{-n},abcdq^{n-1};q)_k q^k}{(q,ab,ac,ad;q)_k}\int_{-1}^{1}w(t;aq^k,b,c,d)\frac{h(t;qz)}{h(t;z)}\,dt. \end{aligned} \tag{2.2}$$

By [7, eq. (3.6)]

$$\begin{aligned} &\int_{-1}^{1}w(t;aq^k,b,c,d)\frac{h(t;qz)}{h(t;z)}\,dt \\ &=\kappa(a,b,c,d)\frac{(bcdz,bzq,czq,dzq;q)_\infty}{(bcdzq,bz,cz,dz;q)_\infty}\frac{(ab,ac,ad;q)_k}{(abcd;q)_k} \\ &\quad\cdot{}_8\phi_7\left[\begin{matrix}bcdz,q\sqrt{\ },-q\sqrt{\ },bc,bd,cd,q,zq^{1-k}/a\\ \sqrt{\ },-\sqrt{\ },dzq,czq,bzq,bcdz,abcdq^k\end{matrix};q,azq^k\right], \end{aligned} \tag{2.3}$$

where the symbol $\sqrt{\ }$ indicates a square root over the top left-hand parameter which, in this case, is $bcdz$. After some simplifications and an easily justifiable interchange in the order of summations, (2.2) and (2.3) give

$$q_n(x)=2z\kappa(a,b,c,d)\frac{(bcdz,bzq,czq,dzq;q)_\infty}{(bcdzq,bz,cz,dz;q)_\infty}\sum_{l=0}^{\infty}\left(\frac{1-bcdzq^{2l}}{1-bcdz}\right)$$
$$\cdot\frac{(bc,bd,cd,zq/a;q)_l}{(dzq,czq,bzq,abcd;q)_l}(az)^l\,{}_3\phi_2\left[\begin{matrix}q^{-n},abcdq^{n-1},a/z\\ abcdq^l,aq^{-l}/z\end{matrix};q,q\right]. \tag{2.4}$$

The ${}_3\phi_2$ series on the right is balanced and terminating and so is summable by Jackson's formula [3, (8.4)] with sum

$$\frac{(q^{1-n+l},bcdzq^l;q)}{(abcdq^l,zq^{1-n+l}/a;q)_n}$$

which vanishes unless $l\geqq n$. Replacing l by $l+n$ in (2.4) and simplifying, we obtain

$$q_n(x)=2z\kappa(a,b,c,d)\frac{(bzq,czq,dzq;q)_\infty}{(bz,cz,dz;q)_\infty}\frac{(bc,bd,cd;q)_n}{(bzq,czq,dzq;q)_n}(az)^n$$
$$\cdot\frac{(q;q)_n(bcdzq^n;q)_{n+1}}{(abcd;q)_{2n}}$$
$$\cdot{}_8\phi_7\left[\begin{matrix}bcdzq,q\sqrt{\ },-q\sqrt{\ },bdq^n,cdq^n,q^{n+1},bcq^n,zq/a\\ \sqrt{\ },-\sqrt{\ },czq^{n+1},bzq^{n+1},bcdzq^n,dzq^{n+1},abcdq^{2n}\end{matrix};q,za\right]. \tag{2.5}$$

Using Bailey's two-term transformation formula [4, (4.3)] for a very well-poised ${}_8\phi_7$ series, we transform the ${}_8\phi_7$ above to the following:

$$\frac{(bcdzq^{2n+1},qz^2,adq^n,abczq^n;q)_\infty}{(az,dzq^{n+1},abcdq^{2n},bcz^2q^{n+1};q)_\infty}$$
$$\cdot{}_8\phi_7\left[\begin{matrix}bcz^2q^n,q\sqrt{\ },-q\sqrt{\ },bz,cz,zq/a,zq/d,bcq^n\\ \sqrt{\ },-\sqrt{\ },czq^{n+1},bzq^{n+1},abczq^n,bcdzq^n,qz^2\end{matrix};q,adq^n\right].$$

Substituting this in (2.5) and simplifying the coefficients, we get

$$q_n(x)=\frac{4\pi z(qz^2;q)_\infty}{(ab,ac,ad,az,bz,cz,dz;q)_\infty}\frac{(abczq^n,bcdzq^n,bzq^{n+1},czq^{n+1},adq^n;q)_\infty}{(bcq^n,bdq^n,cdq^n,q^{n+1},bcz^2q^{n+1};q)_\infty}(az)^n$$
$$\cdot{}_8\phi_7\left[\begin{matrix}bcz^2q^n,q\sqrt{\ },-q\sqrt{\ },bz,cz,zq/a,zq/d,,bcq^n\\ \sqrt{\ },-\sqrt{\ },czq^{n+1},bcq^{n+1},abczq^n,bcdzq^n,qz^2\end{matrix};q,adq^n\right]. \tag{2.6}$$

We now define the q-Wilson function of the second kind:

$$Q_n(z;a,b,c,d)=\frac{1-z^2}{4\pi z}\frac{(az,bz,cz,dz,a/z,b/z,c/z,d/z;q)_\infty}{(z^2,z^{-2};q)_\infty}$$
$$\cdot\frac{(ab,ac,ad;q)_n}{(bc,bd,cd;q)_n}a^{-n}q_n(x) \tag{2.7}$$

$$= \frac{(abczq^n, bcdzq^n, bzq^{n+1}, czq^{n+1}; q)_\infty (a/z, b/z, c/z, d/z; q)_\infty}{(bc, bd, cd, abq^n, acq^n, q^{n+1}, bcz^2q^{n+1}; q)_\infty (z^{-2}; q)_\infty} z^n$$

$$\cdot {}_8\phi_7\left[\begin{matrix} bcz^2q^n, q\sqrt{\ }, -q\sqrt{\ }, bz, cz, zq/a, zq/d, bcq^n \\ \sqrt{\ }, -\sqrt{\ }, czq^{n+1}, bzq^{n+1}, abczq^n, bcdzq^n, qz^2 \end{matrix}; q, adq^n\right].$$

By (1.11), the ${}_8\phi_7$ series is convergent for all z and so the second expression on the right side of (2.7) defines $Q_n(z; a, b, c, d)$ for all z, excluding the poles. The factors in front of $q_n(x)$ in the first expression may appear a bit mysterious at this stage, but the eventual simplifications and symmetries will justify their use. It can be verified through a set of lengthy but straightforward calculations that $Q_n(z; a, b, c, d)$ satisfies the second order divided difference equation of Askey and Wilson [1, (5.16)], implying that it is the right q-analogue of the Jacobi function of the second kind.

Now, by Bailey's formula [4, (5.1)],

(2.8)

$$ {}_8\phi_7\left[\begin{matrix} bcz^2q^n, q\sqrt{\ }, -q\sqrt{\ }, bz, cz, zq/a, zq/d, bcq^n \\ \sqrt{\ }, -\sqrt{\ }, czq^{n+1}, bzq^{n+1}, abczq^n, bcdzq^n, qz^2 \end{matrix}; q, adq^n\right]$$

$$= \frac{(bcz^2q^{n+1}, bdq^n, cdq^n, q^{n+1}, z^{-2}, ab, ac, bc; q)_\infty}{(abcz, bcdzq^n, bzq^{n+1}, czq^{n+1}, b/z, c/z, a/z, dq^n/z; q)_\infty}$$

$$\cdot {}_8\phi_7\left[\begin{matrix} abczq^{-1}, q\sqrt{\ }, -q\sqrt{\ }, az, bz, cz, abcdq^{n-1}, q^{-n} \\ \sqrt{\ }, -\sqrt{\ }, bc, ac, ab, q^{1-n}z/d, abczq^n \end{matrix}; q, q/dz\right]$$

$$- \frac{(bcz^2q^{n+1}, abcq^n/z, bcdq^n/z, bq^{n+1}/z, cq^{n+1}/z, dzq^n, q^{1-n}/dz; q)_\infty}{(bcq^{n+1}/z^2, abczq^n, bcdzq^n, bzq^{n+1}, czq^{n+1}, dq^n/z, zq^{1-n}/d; q)_\infty}$$

$$\cdot \frac{(z^{-2}, qz/d, az, bz, cz; q)_\infty}{(z^2, q/dz, a/z, b/z, c/z; q)_\infty}$$

$$\cdot {}_8\phi_7\left[\begin{matrix} bcq^n z^{-2}, q\sqrt{\ }, -q\sqrt{\ }, b/z, c/z, q/az, q/dz, bcq^n \\ \sqrt{\ }, -\sqrt{\ }, cq^{n+1}/z, bq^{n+1}/z, abcq^n/z, bcdq^n/z, qz^{-2} \end{matrix}; q, adq^n\right]$$

and, by Watson's formula [3, (8.5)],

(2.9)
$$ {}_8\phi_7\left[\begin{matrix} abczq^{-1}, q\sqrt{\ }, -q\sqrt{\ }, az, bz, cz, abcdq^{n-1}, q^{-n} \\ \sqrt{\ }, -\sqrt{\ }, bc, ac, ab, q^{1-n}z/d, abczq^n \end{matrix}; q, q/dz\right]$$

$$= \frac{(abcz, q^{1-n}/ad; q)_n}{(bc, q^{1-n}z/d; a)_n} {}_4\phi_3\left[\begin{matrix} q^{-n}, abcdq^{n-1}, az, a/z \\ ab, ac, ad \end{matrix}; q, q\right]$$

$$= \frac{(abcz, ad; q)_n}{(bc, d/z; q)_n} (az)^{-n} p_n(x; a, b, c, d).$$

Using (2.8) and (2.9) in (2.7), we get

(2.10)
$$Q_n(z; a, b, c, d) = \frac{(ab, ac, ad; q)_n}{(bc, bd, cd; q)_n} a^{-n} p_n(x; a, b, c, d) - Q_n(z^{-1}; a, b, c, d).$$

The symmetry of this formula implies that this is valid for both $|z|<1$, and $|z|>1$. Use of (2.7) in (1.12) and (1.13) now yields

$$q_n(x;a,b,c,d)=\pi i w(x)\left[Q_n(e^{i\theta};a,b,c,d)-Q_n(e^{-i\theta};a,b,c,d)\right], \tag{2.11}$$

$$p_n(x;a,b,c,d)=\frac{(bc,bd,cd;q)_n}{(ab,ac,ad;q)_n}a^n\left[Q_n(e^{i\theta};a,b,c,d)+Q_n(e^{-i\theta};a,b,c,d)\right] \tag{2.12}$$

where $x=\cos\theta$, $0<\theta<\pi$, and $q_n(x;a,b,c,d)$ is the q-Wilson function of the second kind on $-1<x<1$. Note that (2.10) is just the analytic continuation of (2.12) and so it was not really necessary to use (2.8) and (2.9). However, it is interesting to see that Bailey's three-term transformation formula [4, (5.1)] for a very-well-poised ${}_8\phi_7$ series produces the same result.

Askey, Koornwinder and Rahman [2] used a special case of $q_n(x;a,b,c,d)$ by defining the q-ultraspherical functions of the second kind as follows:

$$D_n(\cos\theta;\beta|q)=\frac{4}{w_\beta(\cos\theta|q)}\sum_{k=0}^{\infty}b(k,n;\beta)\cos(n+2k+1)\theta, \tag{2.13}$$

where

$$w_\beta(\cos\theta|q)=\csc\theta\prod_{n=0}^{\infty}\frac{1-2(2\cos^2\theta-1)q^n+q^{2n}}{1-2(2\cos^2\theta-1)\beta q^n+\beta^2q^{2n}}, \tag{2.14}$$

$$b(k,n;\beta)=\frac{(\beta,\beta q;q)_\infty}{(q,\beta^2;q)_\infty}\frac{(\beta^2;q)_n(q\beta^{-1};q)_k(q;q)_{n+k}}{(q;q)_n(q;q)_k(\beta q;q)_{n+k}}\beta^k. \tag{2.15}$$

Since there is no apparent similarity between (2.13) and the expression on the right of (2.7), it is of interest to see how (2.7) and (2.11) give rise to (2.13) in the ultraspherical case.

Let us set $z=e^{i\theta}$, $a=-d=\sqrt{q}$ and $b=-c=\beta$ in (2.7) to get

(2.16)

$$Q_n(e^{i\theta};\sqrt{q},\beta,-\beta,-\sqrt{q})=\frac{(-\beta^2e^{i\theta}q^{n+1/2},\beta^2e^{i\theta}q^{n+1/2},\beta e^{i\theta}q^{n+1},-\beta e^{i\theta}q^{n+1};q)_\infty}{(\beta q^{n+1/2},-\beta q^{n+1/2},q^{n+1},-\beta^2e^{2i\theta}q^{n+1};q)_\infty}$$

$$\cdot\frac{(\sqrt{q}e^{-i\theta},-\sqrt{q}e^{-i\theta},\beta e^{-i\theta},-\beta e^{-i\theta};q)_\infty}{(-\beta^2,-\beta\sqrt{q},\beta\sqrt{q},e^{-2i\theta};q)_\infty}$$

$$\cdot e^{in\theta}\cdot{}_8\phi_7\left[\begin{matrix}-\beta^2e^{2i\theta}q^n,q\sqrt{\ },-q\sqrt{\ },\beta e^{i\theta},-\beta e^{i\theta},\sqrt{q}e^{i\theta},-\sqrt{q}e^{i\theta},-\beta^2q^n\\ \sqrt{\ },-\sqrt{\ },-\beta e^{i\theta}q^{n+1},\beta e^{i\theta}q^{n+1},-\beta^2e^{i\theta}q^{n+1/2},\beta^2e^{i\theta}q^{n+1/2},qe^{2i\theta}\end{matrix};q,-q\right]$$

$$=\frac{(\beta e^{i\theta}q^{n+1},-\beta e^{i\theta}q^{n+1},\beta^2q^n,e^{i\theta}q^{n+3/2},\sqrt{q}e^{-i\theta},-\sqrt{q}e^{-i\theta},\beta e^{-i\theta},-\beta e^{-i\theta};q)_\infty}{(-\beta^2,\beta\sqrt{q},-\beta\sqrt{q},\beta q^{n+1/2},-\beta q^{n+1/2},q^{n+1},-q^{n+1},q^{n+2}e^{2i\theta},e^{-2i\theta};q)_\infty}$$

$$\cdot e^{in\theta}\cdot{}_8\phi_7\left[\begin{matrix}q^{n+1}e^{2i\theta},q\sqrt{\ },-q\sqrt{\ },qe^{i\theta}/\beta,-qe^{i\theta}/\beta,q^{n+1},\sqrt{q}e^{i\theta},-\sqrt{q}e^{i\theta}\\ \sqrt{\ },-\sqrt{\ },\beta e^{i\theta}q^{n+1},-\beta e^{i\theta}q^{n+1},qe^{2i\theta},e^{i\theta}q^{n+3/2},-e^{i\theta}q^{n+3/2}\end{matrix};q,\beta^2q^n\right],$$

by [4, (4.3)]. Now we shall need to use the following quadratic transformation formula recently found by Gasper and Rahman [5].

$$(2.17)\quad {}_8\phi_7\left[\begin{matrix} ax^2/b^2, q\sqrt{\ }, -q\sqrt{\ }, x, -x, x\sqrt{q}/b, -x\sqrt{q}/b, a \\ \sqrt{\ }, -\sqrt{\ }, axq/b^2, -axq/b^2, ax\sqrt{q}/b, -ax\sqrt{q}/b, qx^2/b^2 \end{matrix}; q, aq/b^2\right]$$
$$= \frac{(q/b^2, aqx^2/b^2; q)_\infty}{(aq/b^2, qx^2/b^2; q)_\infty} \frac{(qx^2/b^2, q^2a^2/b^2, qa^2/b^2, q^2x^2/b^4; q^2)_\infty}{(q/b^2, q^2/b^2, qa^2x^2/b^2, q^2a^2x^2/b^4; q^2)_\infty}$$
$$\cdot {}_2\phi_1\left[\begin{matrix} a^2, b^2 \\ a^2q^2/b^2 \end{matrix}; q^2, q^2x^2/b^4\right],$$

provided $|aq/b^2|<1$, $|qx/b^2|<1$. Setting $a=q^{n+1}$, $b=q/\beta$, $x=qe^{i\theta}/\beta$ in (2.17) and substituting in (2.16), we get

$$(2.18)\quad Q_n(e^{i\theta}; \sqrt{q}, \beta, -\beta, -\sqrt{q})$$
$$= -2i\sin\theta \frac{(\beta^2, \beta^2q^{2n+2}, \beta^2e^{2i\theta}, \beta^2e^{-2i\theta}; q^2)_\infty}{(\beta^4, q^{2n+2}, e^{2i\theta}, e^{-2i\theta}; q^2)} e^{i(n+1)\theta}$$
$$\cdot {}_2\phi_1\left[\begin{matrix} q^{2n+2}, q^2/\beta^2 \\ \beta^2q^{2n+2} \end{matrix}; q^2, \beta^2e^{2i\theta}\right].$$

In deriving this we have made frequent use of the identities

$$(2.19)\quad (a;q)_\infty = (a;q^2)_\infty (aq;q^2)_\infty,$$
$$(a^2;q^2)_\infty = (a;q)_\infty (-a;q)_\infty.$$

From (2.11) and (2.18) we then have

(2.20)

$$q_n(x; \sqrt{q}, \beta, -\beta, -\sqrt{q})$$
$$= 4\pi \frac{(\beta^2, \beta^2q^{2n+2}; q^2)_\infty}{(\beta^4, q^{2n+2}; q^2)_\infty} \sum_{k=0}^{\infty} \frac{(q^{2n+2}, q^2/\beta^2; q^2)_k}{(q^2, \beta^2q^{2n+2}; q^2)_k} \beta^{2k} \cos(n+1+2k)\theta$$
$$= 4\pi \frac{(\beta^2, \beta^2q^2; q^2)_\infty}{(q^2, \beta^4; q^2)_\infty} \sum_{k=0}^{\infty} \frac{(q^2\beta^{-2}; q^2)_k (q^2; q^2)_{n+k}}{(q^2; q^2)_k (\beta^2q^2; q^2)_{n+k}} \beta^{2k} \cos(n+1+2k)\theta$$
$$= \frac{(q^2; q^2)_n}{(\beta^4; q^2)_n} \pi w_{\beta^2}(\cos\theta | q^2) D_n(\cos\theta; \beta^2 | q^2).$$

3. Asymptotic properties. Following Ismail and Wilson [6] let us denote

$$(3.1)\quad A(z) = (az, bz, cz, dz; q)/(z^2; q)_\infty.$$

Then it is clear from (2.7) that

$$(3.2)\quad Q_n(z; a, b, c, d) \sim z^n A(z^{-1})/(bc, bd, cd; q)_n,$$

as $n \to \infty$, uniformly for x, a, b, c, d in compact sets avoiding the poles $z^2 = q^{-k}$, $k = 0, 1, 2, \cdots$. Using (2.12) we get

$$(3.3) \qquad a^{-n}(ab, ac, ad; q)_n p_n(x; a, b, c, d)$$
$$\sim z^n A(z^{-1}) + z^{-n} A(z) = 2|A(e^{i\theta})|\cos(n\theta - \phi), \qquad n \to \infty,$$

where $\phi = \arg A(e^{i\theta})$, $0 < \theta < \pi$, $|q| < 1$. This agrees with (1.13) of [6] for $x \neq \pm 1$ and inside the ellipse with foci ± 1 and vertices $\pm 1/2[|q|^{1/2} + |q|^{-1/2}]$. By (3.2) and (2.11) we also get the leading term in the asymptotic expansion of the q-Wilson function of the second kind:

$$(3.4) \qquad (bc, bd, cd; q)_n q_n(x; a, b, c, d) \sim -2\pi w(x)|A(e^{i\theta})|\sin(n\theta - \phi),$$

as $n \to \infty$ for x restricted as above.

REFERENCES

[1] R. Askey and J. A. Wilson, *Some basic hypergeometric orthogonal polynomials that generalize Jacobi polynomials*, Mem. Amer. Math. Soc., 319, Vol. 54, 1985.

[2] R. Askey, Tom H. Koornwinder and Mizan Rahman, *An integral of products of ultraspherical functions and a q-extension*, Proc. Lond. Math. Soc., to appear.

[3] W. N. Bailey, *Generalized Hypergeometric Series*, Stechert–Hafner Service Agency, New York and London, 1964.

[4] _____, *Series of hypergeometric type which are infinite in both directions*, Quart. J. Math. (Oxford), 7 (1936), pp. 105–115.

[5] G. Gasper and Mizan Rahman, *Positivity of the Poisson kernel for the continuous q-Jacobi polynomials and some quadratic transformation formulas for basic hypergeometric series*, to appear.

[6] M. E. H. Ismail and J. A. Wilson, *Asymptotic and generating relations for the q-Jacobi and ${}_4\phi_3$ polynomials*, J. Approx. Theory, 36 (1982), pp. 43–54.

[7] B. Nassrallah and Mizan Rahman, *Projection formulas, a reproducing kernel and a generating function for q-Wilson polynomials*, this Journal, 16 (1985), pp. 186–197.

SIAM J. MATH. ANAL.
Vol. 17, No. 5, September 1986

017

ERRATUM AND ADDENDUM: ANALYTIC FUNCTIONS RELATED TO THE DISTRIBUTIONS OF EXPONENTIAL GROWTH*

RICHARD D. CARMICHAEL†

Abstract. Analytic functions in tubes which were previously shown to have distributional boundary values on the distinguished boundary of the tube are now shown to have boundary values on the topological boundary of the tube as well.

Key words. analytic functions in tubes, distributions of exponential growth, distributional boundary value

AMS(MOS) subject classifications. Primary 32A07, 32A10, 32A40, 46F20

1. Correction. In the final printing of [1, p. 1063] the lines between equations (8.8) and (8.9), which are correct as printed, were printed again after (8.9) to the end of the paragraph. The lines in [1, p. 1063] directly after (8.9) to the end of the paragraph should be as follows:

with this Fourier transform being in the L^2 sense. But $m>0$ is arbitrary in (8.9) and independent of the arbitrary compact subcone $C'\subset C$. Thus it follows that (8.9) holds for $z\in T^{C'}$, C' being an arbitrary compact subcone of C, since for arbitrary $z\in T^{C'}$ we can choose $m>0$ such that $z\in T(C';\ m)$. Since we now know (8.9) holds for $z\in T^{C'}$, then (8.1) follows immediately from this. Now that we have (8.1), the conclusions (8.2) and (8.3) follow by exactly the same type of analysis used in the proof of Theorem 7.1 to prove (7.3) and (7.4).

The remainder of [1, p. 1063] is correct as printed beginning with the paragraph which starts at line 11 ↑.

2. Addition. Let C be an open connected cone in $\mathbb{R}^n$ and let C' be an arbitrary compact subcone of C. If a function $f(z)$ is analytic in $T^{C'}=\mathbb{R}^n+iC'$ for every compact subcone C' of C then $f(z)$ is analytic in the whole of $T^C=\mathbb{R}^n+iC$; for if $z=x+iy\in T^C$, there exists a compact subcone C' of C such that $z\in T^{C'}$ with $y=\mathrm{Im}(z)$ on the interior of C' since C is open. Thus an equivalent definition of the functions $F_p(A;\ C)$ of [1, p. 1053] is that $f(z)\in F_p(A;\ C)$ if and only if $f(z)$ is analytic in T^C and satisfies [1, (6.4), p. 1053] for every compact subcone C' of C and for all $m>0$. In [1, Thm. 8.1, p. 1062] we have [1, (8.1) and (8.2)] holding for $z=x+iy\in T^{C'}$, $C'\subset C$. But the n-tuple α of nonnegative integers, the function $g(t)$, and the distribution $V=D_t^{\alpha}(g(t))\in\mathscr{K}_1'$ in the proof of [1, Thm. 8.1] are all independent of compact subcones $C'\subset C$ (and of the arbitrary $m>0$) as noted in [1, p. 1063, lines 22–23 ↓]. We have that the right side of both [1, (8.1) and (8.2)] are well defined for all $y\in C$. Thus [1, (8.1) and (8.2)] holding for $z\in T^{C'}$ for every compact subcone $C'\subset C$ implies that

*Received by the editors March 18, 1985. This material is based upon work supported by the National Science Foundation under grant DMS-8418435.

† Department of Mathematical Sciences, New Mexico State University, Las Cruces, New Mexico 88003. Permanent address: Department of Mathematics and Computer Science, Wake Forest University, Winston-Salem, North Carolina 27109.

[1, (8.1) and (8.2)] hold for all $z \in T^C$ since $f(z)$ is actually analytic in the whole of T^C here. Additionally [1, (8.3)] holds for $y = \mathrm{Im}(z) \in C$ with $|y| \leqq Q$.

Now let y_0 denote any point in $\mathbb{R}^n$. An easy adaptation of the proof of [1, Lemma 5.9, p. 1052] yields that if $V \in \mathscr{K}_r'$, $r \geqq 1$, then

$$\lim_{\substack{y \to y_0 \\ y \in \mathbb{R}^n}} \exp(-2\pi\langle y, t\rangle) V_t = \exp(-2\pi\langle y_0, t\rangle) V_t \tag{1}$$

in the strong (and weak) topology of $\mathscr{K}_r'$. ($V \in \mathscr{K}_r'$, $r \geqq 1$, implies $(\exp(-2\pi\langle y, t\rangle)V_t) \in \mathscr{K}_r'$ for any $y \in \mathbb{R}^n$ because $\exp(-2\pi\langle y, t\rangle)$ is a multiplier in $\mathscr{K}_r$, $r \geqq 1$, as a function of $t \in \mathbb{R}^n$.) But the Fourier transform defined by [1, (4.2), p. 1044] is a strongly continuous mapping from $\mathscr{K}_r'$ onto K_r'; thus by (1) and the fact that [1, (8.2)] actually holds for all $z = x + iy \in T^C$, as noted in the preceding paragraph, we have

$$\lim_{\substack{y \to y_0 \\ y \in C}} f(x+iy) = \mathscr{F}\left[\exp(-2\pi\langle y_0, t\rangle) V_t\right] \in K_r', \quad r \geqq 1, \tag{2}$$

in the strong (and weak) topology of K_r', $r \geqq 1$, for y_0 being any point on the topological boundary of the cone C. The limit (2) is obtained unrestrictedly, that is independently of how $y \to y_0$, $y \in C$, and is also obtained uniquely. Thus the boundary value conclusion [1, (8.4)] can be replaced by the more general boundary value conclusion obtained in (2) above which yields that the functions $f(z) \in F_1(A;\ C)$, $A \geqq 0$, considered in [1, Theorem 8.1] obtain strong K_r', $r \geqq 1$, boundary values on the topological boundary of the tube T^C as well as on the distinguished boundary $\{x + iy : x \in \mathbb{R}^n, y = \bar{0}\}$ of the tube. (If $y_0 = \bar{0}$, the origin in $\mathbb{R}^n$, (2) reduces to [1, (8.4)].)

The conclusions in [1, Thms. 8.2, 8.4, 9.1, and 9.2] correrssponding to [1, (8.1), (8.2), and (8.3)], where applicable, also hold for $z \in T^C$ ($\in T^{O(C)}$), and [1, (8.10) and (9.2)] hold for all $z \in T^C$. In each of these theorems the boundary value result can be replaced by the more general conclusion (2) for y_0 being any point on the topological boundary of C ($O(C)$).

REFERENCE

[1] R. D. CARMICHAEL, *Analytic functions related to the distributions of exponential growth*, this Journal, 10 (1979), pp. 1041–1068.

SIAM J. MATH. ANAL.
Vol. 17, No. 6, November 1986

PERIODIC SOLUTIONS OF PERIODIC COMPETITIVE AND COOPERATIVE SYSTEMS*

HAL L. SMITH†

Abstract. The periodic solutions, their basins of attraction and invariant manifolds are considered for periodic systems of differential equations which are cooperative or competitive following Hirsch. Competitive and cooperative mappings are introduced which possess the essential features of the Poincaré map for such systems. The geometrical properties of these mappings and the discrete dynamical system they generate are the objects of study. The main tools in this study are the Perron–Frobenius theory of positive matrices and invariant manifold theory. A complete description of the "phase portrait" of the discrete dynamical system generated by an orientation preserving planar cooperative map is obtained.

Key words. periodic solutions, Poincaré map, order preserving diffeomorphism, invariant manifold

AMS(MOS) subject classifications. Primary 34C25, 92A15

Introduction. Many mathematical models in the biological sciences give rise to systems of differential equations

$$\begin{aligned} x_i' &= F_i(x,t), \qquad 1 \leqq i \leqq n, \\ x &= (x_1, x_2, \cdots, x_n) \end{aligned} \tag{0.1}$$

which have quite special properties. Two types of systems which have recently received considerable attention in the literature [10], [11], [12], [15] will be referred to as cooperative and competitive following Hirsch [5]. System (0.1) is cooperative if

$$\frac{\partial F_i}{\partial x_j} \geqq 0, \qquad i \neq j$$

and competitive if the reverse inequalities hold. Applied mathematicians working with particular cooperative or competitive autonomous systems typically found that the asymptotic behavior consisted of convergence to equilibrium (particularly for cooperative systems, e.g., [11] but see [15] for a counterexample in the competitive case). However, any thoughts that competitive and cooperative systems might be immune to the plague of ever more complex and chaotic attractors which the dynamical systems people have discovered had to be scrapped following the note of Smale [12]. Smale showed that any vector field on the standard $(n-1)$-simplex in R^n can be embedded in a smooth competitive vector field on R^n for which the simplex is an attractor. Discouraging as this result might seem for proving any general results for competitive or cooperative systems, it suggests that these systems can behave no worse than general systems in one fewer dimension. In essence, this is what Hirsch showed in [5], [6]. In this series of important papers, many key ideas were introduced.

The above-mentioned work focuses on autonomous systems. We are interested in (0.1) when F_i is periodic in the time variable (of normalized period 2π). Periodic systems (0.1) arise naturally in population biology when day-night cycles or seasonal variation in parameters are accounted for.

*Received by the editors September 18, 1984, and in revised form July 8, 1985.
† Department of Mathematics, Arizona State University, Tempe, Arizona 85287.

Autonomous systems are, of course, periodic and hence all the limitations of Smale's result on any general theory for periodic competitive or cooperative systems are present. Indeed, one expects periodic systems to be less well behaved since they may be viewed as an autonomous system in one more dimension, $S^1 \times R^n$. The first important results for periodic competitive systems were obtained by de Mottoni and Schiaffino [10] who considered periodic two-dimensional Lotka–Volterra systems. They translated properties of competitive planar systems (0.1) to properties of the associated Poincaré map, T, defined by: $x(0) \xrightarrow{T} x(2\pi)$. The most important properties are that T^{-1} preserves the usual partial ordering on R^2 and T is orientation preserving. They then analysed the discrete dynamical system $x_{n+1} = Tx_n$, showing that all orbits $\{T^n x\}_{n \geqq 0}$ tend to fixed points of T. In other words, all solutions of (0.1) are asymptotic to 2π-periodic solutions. In addition, they showed that the nontrivial fixed points of T lie on a separatrix curve: a monotone invariant curve for T forming part of the boundary of the domain of replusion of the trivial fixed point. While de Mottoni and Schiaffino considered only the periodic Lotka–Volterra equations, their arguments were quite general, as pointed out by J. Hale and S. Somolinos [3].

The purpose of this paper is to study the periodic solutions of periodic competitive and cooperative systems (0.1), their respective domains of attraction and invariant manifolds. Following [10], our approach will be to study the geometrical properties of the discrete dynamical system generated by the Poincaré map for (0.1). In the language of discrete dynamical systems, the focus of our work is on fixed points and periodic points of what we term competitive or cooperative mappings. In this sense, the scope of our work is much less ambitious than that of Hirsch [5], [6] since we do not consider the more general problem of determining properties of the limit set $\Lambda(x) = \{y : y = \lim_{i \to \infty} T^{n_i} x$, $\{T^{n_i} x\}$ a convergent subsequence of $\{T^n x\}_{n \geqq 0}\}$ of an orbit $\{T^n x\}_{n \geqq 0}$ (see §2, Proposition L for an exception). In future work we intend to consider this much deeper problem.

The organization of this paper and our main results will be briefly described. In §1 we introduce the Poincaré map for competitive and cooperative periodic systems (0.1) and present some more or less known results concerning these maps, the most important of which are derived from results of Kamke (see [2]) and Hirsch [6]. For a cooperative periodic system (0.1), these properties are, in the above order, that T preserves the usual coordinate wise ordering on R^n ($x < y$ if and only if $x_i < y_i$, $1 \leqq i \leqq n$) and, if $D_x F$ is an irreducible matrix, then $D_x T$ is a positive matrix (all entries are positive). If (0.1) is a competitive system, then T^{-1} enjoys the above properties. Proofs are presented for some of the less well-known results for completeness.

In §2, we introduce the class of cooperative mappings. These maps are diffeomorphisms defined on a neighborhood of R^n_+, the nonnegative orthant, which map the interior, $\dot{R}^n_+$, of R^n_+ into itself, preserve the usual coordinate wise ordering and have the property that the Jacobian derivative of the map evaluated at every positive fixed point is strongly positive (some positive power of the matrix has all positive entries). Included in this class of mappings are the Poincaré maps of cooperative periodic systems (0.1) provided $D_x F$ is irreducible. Nontrivial cooperative maps are shown to have a minimal fixed point which, if it lies in $\dot{R}^n_+$ and is hyperbolic, must be asymptotically stable. We examine the geometry of the basin of attraction of stable fixed points x of a cooperative map T obtaining useful results on the intersection of the basin of attraction of x with the two cones $x + R^n_+$ and $x - R^n_+$. One of these results (Proposition 2.3) implies that if the intersection of the basin of attraction with $x + R^n_+$ is bounded then the boundary of this set, relative to $x + R^n_+$, contains at least one fixed point of T.

Generically, these latter fixed points are unstable. Our main contribution in §2 is the application of a result of the author [13, Thm. 2.2] concerning unstable fixed points of order preserving mappings to cooperative maps. The Perron–Frobenius theory of nonnegative matrices [7], [14] plays a fundamental role in this result and others in this work. Essentially, this result asserts that to each unstable fixed point $x_1 > 0$ of a cooperative map T there are two invariant curves $C^+ \subset x_1 + R^n_+$ and $C^- \subset (x_1 - R^n_+) \cap R^n_+$ issuing from x_1 and parametrized by monotone increasing, respectively decreasing, functions of the half line $[0, \infty)$. C^- connects x_1 to another fixed point x_0, $x_0 < x_1$, which is generically stable and C^+ either connects x_1 to infinity or to another fixed point x_2, $x_2 > x_1$, which is generically stable (the latter necessarily occurs if all orbits, $\{T^n x\}_{n \geqq 0}$, are bounded). In addition, x_0 attracts all points in $[x_0, x_1]$ except x_1 and x_2 attracts all points of $[x_1, x_2]$ except x_1 under iteration by T (if $x \leqq y$, $[x,y] = \{z : x \leqq z \leqq y\}$). In other words, an unstable fixed point x_1 of T must lie on the common boundary of the basins of attraction of two distinct stable fixed points x_0 and x_2, $x_0 < x_1 < x_2$. This result, Theorem 2.5, together with earlier mentioned results, indicates that the fixed point set of T together with the invariant curves C^+ and C^- associated with each unstable fixed point form a tree-like structure with the minimal fixed point at the base.

The results of §2, while providing useful information about the fixed point set of a cooperative map, provide an incomplete description of the dynamics generated by cooperative mappings. This is inevitable, of course, in view of Smale's result. For orientation preserving cooperative mappings in two dimensions, however, we can completely describe the "phase portrait" for the discrete dynamical system under the following reasonable conditions: (A) the minimal fixed point is positive (lies in $\dot{R}^2_+$), (B) all fixed points are hyperbolic and (C) all orbits $\{T^n x\}_{n \geqq 0}$ are bounded. This complete description is possible because (i) every orbit tends to a fixed point of T and (ii) the structure of the fixed point set, including basins of attraction of stable fixed points, can be completely described. Both results rely heavily on the Jordan curve theorem. The fact that in two dimensions every bounded orbit of a cooperative map tends to a fixed point was first recognized by de Mottoni and Schiaffino [10] (see also Hale and Somolinos [3]). Although they observed the property for competitive maps, their arguments are easily transformed to the case of cooperative maps. Our ability to completely describe the structure of the fixed point set as well as the basins of attraction of stable fixed points stems from the earlier mentioned result on monotone invariant curves issuing from unstable fixed points together with a description of the stable manifold of a saddle fixed point. The stable manifold of a saddle fixed point x of T lies in the second and fourth quadrant centered at x and can be divided into two pieces, one lying in each quadrant. Each piece emanates from x in such a way that it can be parametrized by a map of an interval of the real line into R^2, the components of which are monotone, one increasing the other decreasing. Each of the pieces of the stable manifold either connects the saddle point to a fixed point (repeller), becomes unbounded, or intersects a coordinate axis. This global result, Theorem 4.5, provides the component parts of the "separatrix" curves forming the common boundary of the basins of attraction of two stable fixed points. Our main result, Theorem 4.8, synthesizes the earlier results and provides the complete description of the fixed point set and basins of attraction. This description goes roughly as follows. The stable fixed points make up a totally ordered finite or infinite set $\{y_k\}$, $0 < y_1 < y_2 < \cdots < y_k < \cdots$. Each interval $[y_k, y_{k+1}]$ contains an odd number of unstable fixed points (saddles and repellers) lying on a smooth curve S^k. The curve S^k can be parametrized by either of

the coordinate variables in terms of the other in a monotone decreasing fashion and S^k separates R^2_+ into two components. The odd number of fixed points lying in $[y_k, y_{k+1}]$ can be ordered $x^k_1, \cdots, x^k_{2n_k+1}$ from left to right on S^k with the odd indexed x^k being saddles and the even indexed x^k being repellers. S^k consists of the union of the stable manifolds of the saddle fixed points x^k_{2l+1}, $1 \leqq l \leqq n_k$, together with the repellets x^k_{2j}, $1 \leqq j \leqq n_k$. Finally, the S^k, $k \geqq 1$ are pairwise nonintersecting and S^k forms the upper boundary of the basin of y_k (see Fig. 4.8).

The above results for orientation preserving planar cooperative maps are presented in §4. In §3, immediately following our results on general cooperative maps in §2, we introduce the class of competitive maps. These are diffeomorphisms defined on a neighborhood of R^n_+ which map $\dot{R}^n_+$ into itself, map the boundary of R^n_+ into itself and whose additional properties are most easily, but somewhat inaccurately, described by stating that its inverse is a cooperative map. In fact the results of §3 are obtained by applying the results of §2 to the inverse of the competitive map. The main difficulty with this approach is that the domain of the inverse of a competitive map may not include R^n_+. The inverse of a competitive map is a cooperative map which maps the boundary of R^n_+ into itself and hence must fix the origin (Proposition 3.1). In addition, one expects that there will be nontrivial fixed points of a competitive map on the boundary of R^n_+ and that these will be important (recall the competitive exclusion principle of population biology). Despite this fact, we consider primarily the positive fixed points of competitive maps. The reason for this is that in applications, competitive systems (0.1) typically have the property that each portion of a subspace of R^n spanned by a proper subset of the standard basis vectors which makes up the boundary of R^n_+ is itself invariant for (0.1). In other words, the Poincaré map for a competitive periodic system (0.1) preserves each of these "faces" of R^n_+ and, when restricted to any particular face, is a competitive map. Such maps we term "competitive in each face" Since a fixed point of a map, which is competitive in each face, that lies on the boundary of R^n_+ must lie in the (relative) interior of some lower dimensional face of R^n_+, our results on positive fixed points of competitive maps will apply to the restriction of the map to the appropriate face.

A positive fixed point x of a competitive map T will be an unstable fixed point of T^{-1} if any part of the spectrum of $DT(x)$ lies interior to the unit circle. This simple observation indicates that the main result of §2 on stable fixed points of cooperative mappings, will apply quite generally to positive fixed points of competitive maps. One of the consequences of the main result of §3 (Theorem 3.5) will be that, if all fixed points are hyperbolic, every positive fixed point x_1 is either a repeller (all eigenvalues of the Jacobian exceed unity in modulus) or there is a repeller x_0 in R^n_+, $x_0 < x_1$ and a monotone invariant curve for the map T joining x_0 and x_1. Moreover every point of $[x_0, x_1]$ except x_1 lies in the domain of repulsion of x_1 (the basin of attraction of x_0 under T^{-1}). An important class of competitive maps consists of the Poincaré maps of competitive periodic systems (0.1) for which the divergence of the time dependent vector field $F = (F_i)$ is nonpositive. For these maps, Liouville's theorem [4] implies that $\det DT(x) \leqq 1$ for every positive fixed point x of T. As a consequence, T cannot have any positive repelling fixed point and hence a positive fixed point x_1 of T can exist only if there exists a repeller x_0, $x_0 < x_1$, lying on the boundary of R^n_+. One can also show in this case that no two positive fixed points can be related (more generally, two positive fixed points can be related only if the smaller is a repeller or a repeller lies between them).

We have not attempted in this paper to describe the complete phase portrait of a planar orientation preserving competitive map. While the results of §§2 and 4 would be

useful in such a description, the complete description does not immediately follow from the analogous description for the cooperative case (Theorem 4.8) because of our assumption there that the minimal fixed point is positive. One of the reasons for not attempting in this work a description of the dynamics of planar orientation preserving competitive maps is that such a description would depend on both the stability properties of the zero fixed point and on the number of nontrivial fixed points on each of the two coordinate axes neither of which are constrained. It would be reasonable from an applications viewpoint to assume the existence of at most one nontrivial fixed point on each coordinate axis. In any case, the description would break down into several cases. One such case has essentially already been treated by de Mottoni and Schiaffino [10] and Hale and Somolinos [3]. In a future paper, we plan to consider the phase portrait of orientation preserving maps which are competitive in each face in both two and three dimensions in much more detail. The results of our work here will provide the necessary tools for this future study.

1. Competitive and cooperative systems: The Poincaré map. Consider the 2π-periodic system of differential equations

$$x' = F(t,x) = F(t+2\pi,x), \qquad x \in R^n \tag{1.1}$$

to be defined for $(t,x) \in R \times U$ where U is some open neighborhood of $R^n_+ = \{x : x_i \geqq 0, 1 \leqq i \leqq n\}$. We assume F is continuous on its domain and F is C^2 in x for each $t \in R$.

Following Hirsch [5] and Hale and Somolinos [3], we call (1.1)

$$\text{(strictly) competitive if } \frac{\partial F_i}{\partial x_j} \leqq 0 \quad \text{for } i \neq j, \quad (<),$$

$$\text{(strictly) cooperative if } \frac{\partial F_i}{\partial x_j} \geqq 0 \quad \text{for } i \neq j, \quad (>).$$

The inequalities are assumed to hold for all $(t,x) \in R \times R^n_+$. Note that if (1.1) is competitive (cooperative) then the time reversed system

$$y' = -F(-t,y) \tag{1.2}$$

which has solutions $y(t) = x(-t)$ where $x(t)$ satisfies (1.1), is cooperative (competitive).

It is convenient to introduce the following notation. For vectors x and y in R^n we write $x \leqq y$ $(x < y)$ if $x_i \leqq y_i$ $(x_i < y_i)$, $1 \leqq i \leqq n$. If $x \leqq y$, let $[x,y]$ denote the set $\{z \in R^n : x \leqq z \leqq y\}$. We let $R^n_+ = \{x \in R^n : x \geqq 0\}$ and $\dot{R}^n_+ = \{x \in R^n : x > 0\}$. We say x and y are weakly related in case $x \leqq y$ or $y \leqq x$, related if $x < y$ or $y < x$, and unrelated if they are not related. If A is an $n \times n$ matrix we write $A \geqq 0$ $(A > 0)$ if every $a_{ij} \geqq 0$ $(a_{ij} > 0)$. Let $\phi(t,s,x)$ be the solution map for (1.1), i.e., $x(t) = \phi(t,s,x)$ is the maximally extended solution of (1.1) satisfying $x(s) = x$.

In the applications, it is typical for cooperative systems to have the property that if $x(0) \in R^n_+$ then $x(t) \in \dot{R}^n_+$ for $t > 0$ but $x(t) \notin R^n_+$ for $t < 0$, that is, typically boundary points of R^n_+ are strict ingress points of R^n_+ (see [4]). On the other hand, it is typical of competitive systems that each coordinate axis and each of the faces of various dimensions making up the boundary of R^n_+ are invariant sets for (1.1). Hence, in so far as solutions exist in both forward and backward time for (1.1), $\dot{R}^n_+$ is invariant for (1.1) in the competitive case but only forward invariant $(t \geqq 0)$ for the cooperative case. We will make this difference a part of our hypotheses.

Standing hypothesis (SH). If (1.1) is competitive, $s \in R$, $x_0 > 0$ then domain $\phi(\cdot, s, x_0)$ contains $[s, \infty)$ and $\phi(t, s, x_0) > 0$ for all t in the domain of ϕ. If (1.1) is cooperative, $s \in R$, $x_0 > 0$, then domain $\phi(\cdot, s, x_0)$ contains $[s, \infty)$ and $\phi(t, s, x_0) > 0$ for all $t \geqq s$.

In particular, we are assuming the solutions of (1.1) in both cases can be extended into the future but no assumptions are made concerning their extendibility into the past.

Our goal is to draw general conclusions about how periodic solutions of (1.1) are situated. It is important, in order to make our conclusions as generally applicable as possible, to restrict our consideration to solutions of (1.1) which lie in $\dot{R}^n_+$. While this restriction may appear to ignore important (and sometimes all) the interesting periodic behavior, we point out that our results can usually be applied to the lower dimensional hyperplanes forming the boundary of R^n_+ (see §3).

The fundamental result for competitive and cooperative systems is due to Kamke which we state as follows [2].

THEOREM A. *Let $x(t)$ and $y(t)$ be solutions of* (1.1) *on* $[a, b]$.

(i) *If* (1.1) *is cooperative and $x(a) < y(a)$ then $x(b) < y(b)$.*

(ii) *If* (1.1) *is competitive and $x(b) < y(b)$ then $x(a) < y(a)$.*

Note that (ii) is equivalent to the corresponding versions in Hirsch [5] or Hale and Somolinos [3]. It follows from Theorem A and continuity of solutions with respect to initial conditions that the strict inequalities in (i) can be replaced by inequalities; a corresponding result may be obtained in (ii). In two dimensions, Theorem A has the following corollary, the proof of which the reader may easily supply.

COROLLARY B. *Let $n = 2$. Let $z(t)$ and $y(t)$ be solutions of* (1.1), $y_i(t) = z_i(t) + h_i(t)$.

(i) *If* (1.1) *is cooperative and $h_1(a)h_2(a) < 0$ then it cannot be the case that $h_1(a)h_1(b) < 0$ and $h_2(a)h_2(b) < 0$.*

(ii) *If* (1.1) *is competitive and $h_i(a) \geqq 0$, $h_j(a) \leqq 0$, then $h_i(b) \geqq 0$ and $h_j(b) \leqq 0$, $\{i, j\} = \{1, 2\}$. If $h_1(a)h_2(a) > 0$ then it cannot be the case that $h_1(a)h_1(b) < 0$ and $h_2(a)h_2(b) < 0$.*

Another corollary of Theorem 1 will be useful in our study of competitive systems.

COROLLARY C. *Let* (1.1) *be competitive, $t_1 \geqq 0$, $x_0 \in R^n_+$ and $x_1 = \phi(t_1, 0, x_0)$. Then* $\phi(t_1, 0, [0, x_0]) \supset [0, x_1]$.

Proof. Let $x \in [0, x_1]$ and let $z(t)$ be the maximal solution of (1.2) satisfying $z(-t_1) = x$. Since $y(t) \equiv \phi(-t, 0, x_0)$, $-t_1 \leqq t \leqq 0$, and $z(t)$ satisfy (1.2) and $y(-t_1) = x_1 \geqq x = z(-t_1)$, Theorem A together with (SH) imply that $z(t)$ is defined on $[-t_1, 0]$ and $y(t) \geqq z(t)$, $-t_1 \leqq t \leqq 0$. One easily shows that $z_0 = z(0) \leqq x_0$ satisfies $\phi(t_1, 0, z_0) = x$.

Note that Corollary C implies that $\phi(t, 0, 0) \equiv 0$.

We now consider the variational equation for (1.1) corresponding to a solution $\phi(t, 0, x)$, $x > 0$, namely

$$X' = D_x F(t, \phi(t, 0, x))\, X, \qquad X(0) = I. \tag{1.3}$$

Let $\Phi(t, x)$ be the (fundamental) matrix solution of (1.3), i.e., $\Phi(t, x) = D_x \phi(t, 0, x)$. The result below appears in Hirsch [6] but a slightly weaker version may be found in Krasnosel'skii [8]. An $n \times n$ matrix $A = (a_{ij})$ is irreducible if whenever the set $\{1, 2, \cdots, n\}$ is expressed as the union of two disjoint proper subsets S, S', then for every $i \in S$ there exists j and k in S', such that $a_{ij} \neq 0$ and $a_{ki} \neq 0$. This means that the linear transformation A does not map into itself any nonzero proper linear subspace spanned by a subset of the standard basis vectors.

THEOREM D. (i) *If* (1.1) *is cooperative* (*competitive*) *then* $\Phi(t,x)\geqq 0$ ($\Phi^{-1}(t,x)\geqq 0$) *for* $t\geqq 0$.

(ii) *If* $D_xF(t,z)$ *is irreducible for all* t *and* $z>0$ *and* (1.1) *is cooperative* (*competitive*) *then* $\Phi(t,x)>0$($\Phi^{-1}(t,x)>0$) *for* $t>0$.

Actually, only the cooperative case appears in Hirsch [6] but the competitive case of Theorem D follows immediately from applying the results of the cooperative case to the adjoint equation corresponding to (1.3), which is cooperative if (1.1) is competitive, thus deducing that the transpose of $\Phi(t,x)^{-1}$ is nonnegative (positive).

It is worth noting that $D_xF(t,z)$ is irreducible if (1.1) is either strictly competitive or strictly cooperative.

We will refer to (1.1) as "cooperative (competitive) and irreducible" if (1.1) is cooperative (competitive) and $D_xF(t,z)$ is irreducible for $t\in R$ and $z>0$. Actually, we do not require the strong result that $\Phi(t,x)>0$ for $t>0$, $x>0$ but only that for each $x>0$, $\Phi(t,x)>0$ for large t. This property is referred to by Hirsch [6] as "$\phi(t,s,x)$ has eventually positive derivatives". Similarly, if (1.1) is competitive we need only require that $\Phi(t,x)^{-1}>0$ for large t. While there are advantages to assuming the weaker condition (the property is inherited by nearby systems), it does not appear to be readily verifiable in applications.

As noted by Hirsch [6], systems (1.1) which are cooperative (competitive) and irreducible possess stronger monotonicity properties.

COROLLARY E. *Let* $x_1\geqq 0$, $x_2>0$ *be distinct and* $x_1\leqq x_2$.

(i) *If* (1.1) *is cooperative and irreducible, then* $\phi(t,0,x_1)<\phi(t,0,x_2)$ *for* $t>0$.

(ii) *If* (1.1) *is competitive and irreducible, then* $\phi(0,t,x_1)<\phi(0,t,x_2)$ *for* $t>0$.

Proof. For (i), note that

$$\phi(t,0,x_2)-\phi(t,0,x_1)=\int_0^1\Phi(t,sx_2+(1-s)x_1)\,ds(x_2-x_1).$$

For (ii), note that $x=\phi(0,t,\phi(t,0,x))$ so

$$\phi(0,t,x_2)-\phi(0,t,x_1)=\int_0^1\Phi^{-1}(t,\phi(0,t,sx_2+(1-s)x_1))\,ds(x_2-x_1).$$

In either case, the result follows from Theorem D.

Another property that (1.1) may possess in either the competitive or cooperative case is that

$$\operatorname{div}F(t,x)\leqq 0,\qquad (t,x)\in R\times\dot{R}^n_+ \tag{1.4}$$

where div denotes the divergence of the time-dependent vector field F. In case (1.4) holds, Liouville's theorem [4] states that

$$0<\det\Phi(t,x)\leqq 1. \tag{1.5}$$

We will see that (1.5) has particularly strong consequences for competitive systems.

We are now ready to define the primary object of study in this paper, the Poincaré map. The Poincaré map for (1.1) is defined for $x\in R^n_+$ as follows

$$T(x)=\phi(2\pi,0,x).$$

It is well known that T is a C^2 orientation preserving diffeomorphism defined on a neighborhood of R^n_+ whose fixed points and periodic points (x is a periodic point of T of period p if $T^px=x$ and $T^jx\neq x$ for $1\leqq j\leqq p-1$) correspond to 2π, periodic solutions or $2\pi p$-periodic solutions of (1.1).

Theorem A and Corollary E have the following implication for the Poincaré map T.

COROLLARY F. (i) *If* (1.1) *is cooperative and* $0 \leqq x_1 < x_2$ *then* $T(x_1) < T(x_2)$. *If* (1.1) *is also irreducible, the conclusion holds if* $0 \leqq x_1 \leqq x_2$, $x_1 \neq x_2$ *and* $x_2 > 0$.

(ii) *If* (1.1) *is competitive,* $x_i \geqq 0$, $i = 1, 2$, *and* $T(x_1) < T(x_2)$ *then* $x_1 < x_2$. *If* (1.1) *is also irreducible, the conclusion holds if* $Tx_2 > 0$ *and* $Tx_1 \leqq Tx_2$, $x_1 \neq x_2$.

One can view (ii), which we will do, as asserting that $T^{-1}: T(R^n_+) \to R^n_+$, $T^{-1}(x) = \phi(0, 2\pi, x)$ is monotone in the sense of (i). It will be important to have the following information concerning the domain $T(R^n_+)$, of T^{-1} which follows from Corollary C.

COROLLARY G. *Let* (1.1) *be competitive,* $x_0 \geqq 0$, *and* $x_1 = T(x_0)$. *Then* $T([0, x_0]) \supset [0, x_1]$.

In other words, if $x_1 \in T(R^n_+)$ then $[0, x_1] \subset T(R^n_+)$ and if $x \notin T(R^n_+)$ then $y \notin T(R^n_+)$ if $y \geqq x$.

Theorem D(i) or Corollary F imply that the derivative, $DT(x) = \Phi(2\pi, x)$, of T satisfies $DT(x) \geqq 0$ in the cooperative case and $[DT(x)]^{-1} = D(T^{-1})(Tx) \geqq 0$ in the competitive case.

LEMMA H. *Let* (1.1) *be cooperative* (*competitive*) *and irreducible. If* $\phi(t, 0, x_0)$, $x_0 > 0$, *is a* $2\pi j$*-periodic solution of* (1.1) *for some positive integer* j, *i.e.,* $T^j(x_0) = x_0$, *then*

$$(1.6)\qquad \begin{aligned} &\text{(a)} \quad \left[D(T^j)(x_0)\right] > 0 \quad \textit{if (1.1) is cooperative,} \\ &\text{(b)} \quad \left[D(T^{-j})(x_0)\right] > 0 \quad \textit{if (1.1) is competitive.} \end{aligned}$$

Proof. Assume first that (1.1) is cooperative. Then $T^j(x) = \phi(2j\pi, 0, x)$ so by the chain rule and the fact that $T^j(x_0) = x_0$ we have

$$\Phi(2j\pi, x_0) = D_x \phi(2j\pi, 0, x_0) = D(T^j)(x_0).$$

The matrix on the left is assumed to be positive.

If (1.1) is competitive and $T^j(x_0) = x_0$ for $x_0 > 0$ then it follows that $\phi(t, 0, x_0)$ is defined for all $t \in R$. Hence T^{-j} is defined in a neighborhood of x_0. Since $T^{-j}(x_0) = x_0$ we have

$$\left[D(T^{-j})(x_0)\right] = \left[D(T^j)(x_0)\right]^{-1} = \left[\Phi(2j\pi, x_0)\right]^{-1} > 0.$$

We remark that if in Lemma H, irreducible is replaced by "ϕ has eventually positive derivatives" then 1.6 (a) and (b) are modified only by the need to raise the bracketed matrix to a sufficiently large positive integer power in order to have positivity.

Lemma H has important implications for the stability of $2\pi j$-periodic solutions of (1.1). The primary reason for the usefulness of Lemma H is that it allows the application of the Perron–Frobenius theory of positive matrices [7], [14].

THEOREM I. *If* $A \geqq 0$ *is an* $n \times n$ *matrix for which* $A^p > 0$ *for some positive integer* p *then* $\rho(A)$, *the spectral radius of* A, *is a positive simple eigenvalue of* A *strictly exceeding in modulus all other eigenvalues of* A. *Moreover, there exist* $x > 0$ *such that* $Ax = \rho(A)x$ *and* x *is the unique eigenvector of* A (*up to scalar multiple*) *which lies in* R^n_+.

Recall that a fixed point x_0 of a smooth map T is asymptotically stable if the spectral radius of $DT(x_0)$, $\rho(x_0) \equiv \rho(DT(x_0))$, is less than one and unstable if $\rho(x_0)$ is larger than one. If (1.1) is cooperative and irreducible, then Lemma H implies that the derivative of the Poincaré map T at x_0 satisfies $DT(x_0) > 0$ if $x_0 > 0$ and by Theorem I, $\rho(x_0)$ is a simple eigenvalue of $DT(x_0)$ with corresponding eigenspace spanned by a

positive vector. This property will prove to be important in our study of cooperative maps in §2. If (1.1) is competitive and irreducible, then the derivative of the inverse of the Poincaré map will have the property that its spectral radius is a simple eigenvalue with positive eigenvector.

We have now accumulated the results which will be essential in our study of periodic solutions of competitive and cooperative systems (1.1). Our approach will be to study the Poincaré map T. The cooperative case will be considered in greater detail since the competitive case can be treated by applying the results of the cooperative case to T^{-1}. We do, however, consider an important result for the competitive case where (1.1) satisfies the nonpositive divergence condition (1.4) (see §3). In addition, we will largely restrict our attention to fixed points of T and consider explicitly periodic points of T only briefly. The rationale for this is that the conditions we place on T in the results to follow will apply also to T^j for each positive integer j. Thus, by replacing T in the results of §2, by T^j one obtains results concerning periodic points. Having said this, we make some remarks concerning periodic points of T. First, in dimensions one and two, all periodic points in either the competitive or cooperative case are fixed points, i.e., all periodic solutions whose period is an integer multiple of 2π are in fact 2π-periodic. In one dimension this is trivial, all periodic scalar equations are cooperative and T is strictly increasing. In two dimensions the result is not so trivial. It is contained in results proved by de Mottoni and Schiaffino [10] and in greater generality by Hale and Somolinos [3] (see §4). Our second point is the following.

PROPOSITION J. *Let* (1.1) *be competitive or cooperative, let* x_0 *be a periodic point of period* $j>1$ *for* T (*i.e.* $\phi(t,0,x_0)$ *is* $2j\pi$*-periodic*), *and let* $x_i=T^i(x_0)(x_i(t)=\phi(t+2i\pi,0,x_0))$ $i=0,1,2,\cdots,j-1$. *Then no two distinct points* x_i *and* x_j *can be weakly related* (*no two distinct solutions* $x_i(t)$ *and* $x_j(t)$ *can be weakly related for any* t).

Proof. Consider the cooperative case (the competitive case may be treated by considering T^{-1}). If $x_r\leqq x_s$ for distinct integers $r,s\in\{0,1,\cdots,j-1\}$, say $r<s$, then on applying T^{j-r} to the inequality we obtain $x_0\leqq T^p(x_0)$, $p=s-r\in\{1,2,\cdots,j-1\}$. But then $x_0\leqq T^p(x_0)\leqq T^{2p}(x_0)\leqq\cdots\leqq T^{np}(x_0)\leqq\cdots$ and equality cannot occur in any inequality. On the other hand, the points $T^{np}(x_0)$, $n=1,2,\cdots$ all belong to $\{x_i\}_{i=0}^{j-1}$ so they must repeat, i.e., there exists $n,m,n-m>1$ such that

$$T^{mp}(x_0)\leqq T^{(m+1)p}(x_0)\leqq\cdots\leqq T^{np}(x_0)=T^{mp}(x_0).$$

This implies that

$$T^{mp}(x_0)=T^{(m+1)p}(x_0)=\cdots=T^{np}(x_0),$$

a contradiction.

For dimensions larger than two, cooperative or competitive systems (1.1) can have periodic solutions of minimal period $2\pi j$, $j>1$, as well as far more bizarre attractors. This is a consequence of a result of Smale [12] which states that any smooth vector field on the standard $(n-1)$-simplex in R^n_+ can be extended to a smooth competitive vector field on R^n.

Our final comment on periodic points is to point out that a positive stable 2π-periodic solution of a one-parameter family of cooperative irreducible systems can participate in only two types of bifurcation, namely, it can coalesce with another 2π-periodic solution and disappear at a critical value of the parameter or it can persist, throwing off another 2π-periodic solution as it loses stability at a critical value of the parameter. The other possible bifurcations are forbidden by Lemma H and Theorem I: the largest characteristic multiplier in modulus is necessarily positive.

We close this section with two results which are obvious modifications of similar results of Hirsch [5], [6] for the autonomous case. Let T be the Poincaré map for (1.1). If $x_0 \geqq 0$ we write $O^+(x_0)=\{T^n(x_0): n=0,1,2,\cdots\}$ for the forward orbit and $\Lambda(x_0)$ for the set of limit points of $O^+(x_0)$, that is, $\Lambda(x_0)=\{x: x=\lim_{i\to\infty} T^{n_i}x$, where $\{n_i\}_{i=1}^{\infty}$ is a subsequence of the natural numbers$\}$.

PROPOSITION K. *Let* (1.1) *be cooperative and* $O^+(x_0)$ *be a bounded orbit and suppose* $T^n(x_0) \leqq T^m(x_0)$ *for distinct nonnegative integers* n, m. *Then either* $O^+(x_0)$ *is a periodic orbit or* $\Lambda(x_0)$ *is a periodic orbit.*

Proof. If $T^n(x_0)=T^m(x_0)$, then $O^+(x_0)$ is a periodic orbit. Suppose $T^n(x_0) \leqq T^m(x_0)$ but equality does not hold. Then by Proposition J, $O^+(x_0)$ is not a periodic orbit. We assume $n>m$, i.e., $n=m+p$, the other case is similar. Then $T^p(T^m(x_0)) \leqq T^m(x_0)$ so $T^{lp}(T^m(x_0)) \leqq T^{(l-1)p}(T^m(x_0))$, $l=1,2,\cdots$. Since $O^+(x_0)$ is bounded, $y=\lim_{l\to\infty} T^{lp+m}(x_0)$ exists. By continuity of T, $T^p(y)=y$, and $\lim_{l\to\infty} T^{lp+m+r}(x_0) = T^r(y)$ for $r=1,2,\cdots,p-1$. It follows that $\Lambda(x_0)$ is the periodic orbit $\{T^r(y)\}_{r=0}^{p-1}$.

The following result generalizes Proposition J.

PROPOSITION L. *Let* (1.1) *be cooperative and* $O^+(x_0)$ *be a bounded orbit. Then no two elements of* $\Lambda(x_0)$ *can be related.*

Proof. We may assume $O^+(x_0)$ is not a periodic orbit by Proposition J. Suppose $x,y \in \Lambda(x_0)$ and $x<y$. It follows that there exist two distinct integers n and m such that $T^n(x_0)<T^m(x_0)$. By Proposition K, $\Lambda(x_0)$ is a periodic orbit. But then we have a contradiction to Proposition J.

If (1.1) is cooperative and irreducible, then "related" can be replaced by "weakly related" in Proposition L. This follows since $T\Lambda \subset \Lambda$ and if $x \leqq y$ then $T^n x < T^n y$ for some positive integer n.

2. Fixed points of cooperative maps: Invariant manifolds and basins of attraction. In this section we consider C^2 diffeomorphisms defined on a neighborhood of R_+^n which map $\dot{R}_+^n$ into itself and satisfy

(M) $0 \leqq x < y$ implies $Tx < Ty$, and

(SP) $x>0$ and $T(x)=x$ implies $DT(x)^p>0$ for some positive integer p depending on x.

We will call such maps cooperative. It will be assumed without further mention that all maps T in this section are cooperative but in one of our results we will assume the stronger monotonicity assumption

(SM) $0 \leqq x \leqq y$, $x \neq y$ and $y>0$ implies $T^p(x)<T^p(y)$ for large positive integers p.

Recall that if T is the Poincaré map of a C^2 cooperative system (1.1), then (M) holds. If, in addition, (1.1) is irreducible (or has eventually monotone derivatives) then (SM) and (P) hold for T by Corollary F and Lemma H respectively.

The focus of this section is to describe the manner in which the positive fixed points together with their domains of attraction and invariant manifolds are situated in R_+^n. We begin by observing that, in the interesting case, a cooperative map has a minimal fixed point though it may not be positive.

PROPOSITION 2.1. *Either* $|T^n x| \to \infty$ *as* $n \to \infty$ *for every* $x \geqq 0$ *or* T *has a minimal fixed point* $x_m=\lim_{n\to\infty} T^n 0$, *i.e.*, $Tx=x$ *implies* $x \geqq x_m$. $T^n x \to x_m$ *as* $n \to \infty$ *for all* $x \in [0, x_m]$.

Proof. We have $0 \leqq T0 \leqq T^2 0 \leqq \cdots \leqq T^n 0 \leqq \cdots$. The first alternative occurs if $|T^n 0| \to \infty$ while the latter occurs if $\{T^n 0\}_{n\geqq 0}$ is bounded. If $0 \leqq x \leqq x_m$ then $T^p 0 \leqq T^p x \leqq x_m$ for every p so $T^p x \to x_m$ as $p \to \infty$.

The following result will be useful. It follows immediately from (SM) but we prefer to assume only (M) and (SP).

LEMMA 2.2. *Let* $x_0>0$, $x\geqq 0$ *be distinct points and assume* $Tx_0=x_0$. *If* $x\geqq x_0$ ($x\leqq x_0$) *then* $T^p x>x_0$ ($T^p x<x_0$) *for large positive integers* p.

Proof. Suppose $x\geqq x_0$ and let $h=x-x_0\geqq 0$. Since $T^p(x_0+th)=x_0+tD(T^p)(x_0)h+o(t)$ as $t\to 0$, it follows that $t^{-1}(T^p(x_0+th)-x_0)=(DT(x_0)^p h+O(t))$ for small positive t so by (SP), $t^{-1}(T^p(x_0+th)-x_0)>0$ for small positive t, p as in (SP). Hence $T^p(x_0+th)>x_0$ for small t and $x_0<T^p(x_0+th)\leqq T^p(x_0+h)=T^p(x)$.

Lemma 2.2 implies that two nonnegative fixed points of T, at least one of which is positive, are related if they are weakly related. This observation will be important.

We begin our study of positive fixed points of T by establishing some notation. Let $x_0>0$ be a fixed point of T and let $B(x_0)=\{x\geqq 0: T^n x\to x_0 \text{ as } n\to\infty\}$. Let $B^+(x_0)=B(x_0)\cap(x_0+R^n_+)$ and $B^-(x_0)=B(x_0)\cap[0,x_0]$ denote respectively those points in $B(x_0)$ for which $x\geqq x_0$ and $x\leqq x_0$. Let $S^+(x_0)$ denote the boundary of $B^+(x_0)$ considered as a subset of the space x_0+R^n and $S^-(x_0)$ denote the boundary of $B^-(x_0)$ considered as a subset of the space $[0,x_0]$. We drop x_0 from the notation when no confusion will result. Each of the sets B, B^+, B^-, S^+ and S^- is mapped into itself by T. In the following result, some simple properties are derived concerning the sets B, B^+, B^-, S^+ and S^-. We do not assume x_0 is asymptotically stable but in that case, B contains a neighborhood of x_0.

PROPOSITION 2.3. *Let* $x_0>0$ *be a fixed point of* T. *Then*

(a) *If* $x_i\in B$, $i=1,2$, $x_1\leqq x_2$ *then* $[x_1,x_2]\in B$.

(b) *If* $x_2\in S^+$ *and* $x_0\leqq x_1<x_2$ *then* $x_1\in B^+$. *If* $x_1\in S^-$ *and* $x_1<x_2\leqq x_0$ *then* $x_2\in B^-$.

(c) *If* B^+ *contains the intersection of a ball centered at* x_0 *with* $x_0+R^n_+$ *then either* B^+ *is unbounded or* S^+ *is homeomorphic to the* $n-1$ *simplex and contains a fixed point of* T. *If* B^- *contains the intersection of a ball centered at* x_0 *with* $x_0-R^n_+$ *and if* $\partial R^n_+\cap[0,x_0]$ *is contained in the interior of* $[0,x_0]-B^-$ *relative to* $[0,x_0]$ *then* S^- *is homeomorphic to the* $n-1$ *simplex and contains a fixed point of* T.

Proof. (a) follows from $T^n x_1\leqq T^n x\leqq T^n x_2$, $n=1,2,\cdots$, which holds if $x_1\leqq x\leqq x_2$.

The first assertion in (b) follows from the observation that there exists a ball centered at x_2 with the property that all of its points x satisfy $x>x_1$. But one of these points necessarily lies in B^+ so, by (a), $x_1\in B^+$. The second assertion is proved similarly.

If B^+ is bounded, then S^+ is nonempty. In fact for every $h\geqq 0$, $|h|=1$ there is a unique (by (a)) value, t_h, of $t>0$ such that $x_0+t_h h\in S^+$. It is easy to see that $h\to t_h$ is continuous so S^+ is homeomorphic to the $n-1$ simplex. Since $T(S^+)\subset S^+$, the Brouwer fixed point theorem implies S^+ contains a fixed point of T. The assertion concerning S^- is proved in a similar fashion.

If (SM) holds for T then S^+ and S^- can be continuously parametrized by points of the set obtained by projecting S^+ (S^-) orthogonally along a standard basis vector onto an $(n-1)$-dimensional face of ∂R^n_+. For the statement of the result, we require the following notation. For $i=1,2,\cdots,n$, let $H_i(x_0)=\{x: x_i=(x_0)_i\}$, $H^+(x_0)=H_i(x_0)\cap(x_0+R^n_+)$, $H^-(x_0)=H_i(x_0)\cap(x_0-R^n_+)$ and $P_i: R^n\to H_i(x_0)$ be the orthogonal projection onto $H_i(x_0)$ along the ith standard bases vector e_i (see Fig. 2.1). We drop x_0 from the notation when it is clear to which fixed point we refer. In the following result we assume x_0 is an asymptotically stable fixed point in order to avoid a more lengthy statement. All that is required is that B^+ (B^-) contains the intersection of a ball about x_0 with $x_0+R^n_+$ ($x_0-R^n_+$).

PROPOSITION 2.4. *Let* (SM) *hold and* $x_0>0$ *be an asymptotically stable fixed point of* T. *If* $S^+\neq\phi$ (*i.e.*, $B^+\neq x_0+R^n_+$), *then no two distinct elements of* S^+ *can be weakly related nor can they have the same projection onto* H_i *along* e_i *for* $1\leqq i\leqq n$. *For* $1\leqq i\leqq n$,

$P_i(S^+)$ contains $[x_1, x_2]$ whenever $x_1 \leqq x_2$ belong to $P_i(S^+)$. There exists a continuous function $h_i^+ : P_i(S^+) \to R^+$, strictly decreasing in the sense that $x_1 \leqq x_2$, $x_1 \neq x_2$ implies $h_i^+(x_1) > h_i^+(x_2)$, such that $S^+ = \{x + h_i^+(x)e_i : x \in P_i(S^+)\}$.

If $\partial R_+^n \cap [0, x_0]$ is contained in the interior of $[0, x_0] - B^-$ relative to $[0, x_0]$ then no two distinct elements of S^- can be weakly related nor can they have the same projection onto H_i along e_i for $1 \leqq i \leqq n$. For $1 \leqq i \leqq n$, $P(S^-)$ contains $[x_1, x_2]$ whenever $x_1 \leqq x_2$ belong to $P_i(S^-)$. There exists a continuous function $h_i^- : P_i(S^-) \to R^-$, strictly decreasing in the sense that $x_1 \leqq x_2$, $x_1 \neq x_2$ implies $h_i^-(x_2) < h_i^-(x_1)$, such that $S^- = \{x + h_i^-(x)e_i : x \in P_i(S^-)\}$.

Proposition 2.4 implies that $P_i : S^+ \to P_i(S^+)$ is a homeomorphism, the inverse of which is the monotone map $x \to x + h_i^+(x)e_i$. Figure 2.1 below depicts the geometry of the sets S^+ and S^- for a two-dimensional competitive map T satisfying (SM).

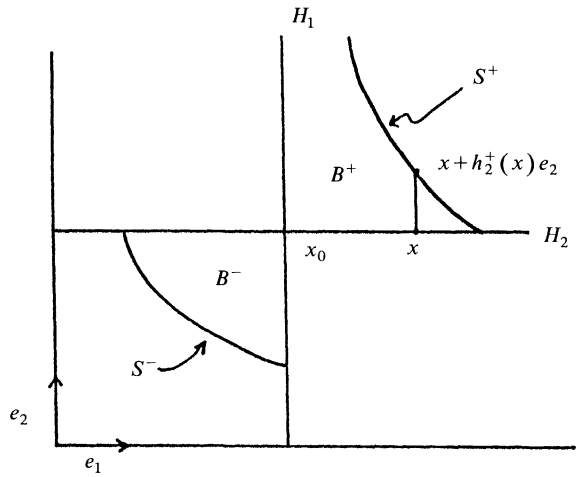


FIG. 2.1. *A possible configuration of the sets B^+, S^+, B^-, S^- for the stable fixed point x_0 of T which is consistent with Proposition* 2.4 *in two dimensions.*

Proof. Suppose $x_1, x_2 \in S^+$, $x_1 \neq x_2$ and $x_1 \leqq x_2$. Then $T^p x_1, T^p x_2 \in S^+$ and $x_0 \leqq T^p x_1 < T^p x_2$ by (SM). This contradicts Proposition 2.3 (b). Note that if $x_2 \in S^+$ and $x_0 \leqq x_1 \leqq x_2$ where x_1 is distinct from x_2 then the above argument shows (by Proposition 2.3 (b)) that $T^p x_1 \in B^+$ for large p and hence $x_1 \in B^+$. Suppose now that x_1 and x_2 are distinct points of S^+ and $P_i x_1 = P_i x_2$ for some i. Then it follows that x_1 and x_2 are related which contradicts the assertion proved above. Let $x_j = P_i y_j$ where $y_j \in S^+$, $j = 1, 2$, are distinct points. Suppose $x_1 \leqq x_2$ and write $y_j = x_j + h_i^+(x_j)e_i$, $j = 1, 2$, where $h_i^+(x_j) \geqq 0$ since $x_j \leqq y_j$, $j = 1, 2$. If $h_i^+(x_1) \leqq h_i^+(x_2)$ then clearly $y_1 \leqq y_2$ contradicting the assertion proved above. Hence $h_i^+(x_1) > h_i^+(x_2)$ proving h_i^+ is strictly decreasing. Note that if $x_2 \neq y_2$ then $x_2 \leqq y_2$ so $x_2 \in B^+$. If $x \in [x_1, x_2]$ is distinct from x_2, then it follows that $x \in B^+$. If $x \notin P_i(S^+)$ then $x + te_i \in B^+$ for all $t > 0$. But $y_1 = x_1 + h_i^+(x_1)e_1 \leqq x + te_1$ for large t, implying that $y_1 \in B^+$ (Proposition 2.3 (a)). This contradiction proves $x \in P_i(S^+)$ whenever $x_1 \leqq x \leqq x_2$ and $x_1, x_2 \in P_i(S^+)$. All that remains to be proved is the continuity of h_i^+. Let $x_n \to x$ as $n \to \infty$ where x_n, $x \in P_i(S^+)$, $n = 1, 2, \cdots$. Let $h_i^+(x_{n_j}) \to \alpha \in [0, \infty)$ as $n_j \to \infty$ where $\{x_{n_j}\}$ is a subsequence of $\{x_n\}$. Since S^+ is closed, $x + \alpha e_i \in S^+$. But then $\alpha = h_i^+(x)$. One sees that $\limsup_{n \to \infty} h_i^+(x_n) < \infty$ as follows. If $t > h_i^+(x)$ then $x + te_i \notin S^+ \cup B^+$. Since $S^+ \cup B^+$ is closed, it follows that $x_n + te_i \notin S^+ \cup B^+$ for all large n. One then must conclude that $h_i^+(x_n) < t$ for these large n. It follows that h_i^+ is continuous. The arguments for S^- are omitted since they parallel those above.

We now turn our attention to fixed points $x>0$ of T satisfying $\rho(x)>1$. Recall, $\rho(x)=\rho(DT(x))$ is the spectral radius of $DT(x)$. Such fixed points are unstable. The following theorem is one of the main results of this paper. In large part, its proof is contained in [13].

THEOREM 2.5. *Let $x_1>0$ be a fixed point of T with $\rho_1=\rho(x_1)>1$ and $DT(x_1)e_1=\rho_1 e_1$ where $e_1>0$. Then there exists a C^1 function $y_+:[0,\infty)\to \dot{R}^n_+$ satisfying*

(A_+) $y_+(t)=x_1+te_1+O(t^2)$ *as* $t\to 0$.

(B_+) $0\leqq t_1<t_2$ *implies* $y_+(t_1)<y_+(t_2)$.

(C_+) $y_+(t)=T(y_+(\rho_1^{-1}t))$, $t\geqq 0$.

(D_+) *Either* $\lim_{t\to\infty}|y_+(t)|=\infty$ *or* $\lim_{t\to\infty}y_+(t)=x_2>x_1$. *In the latter case,* $Tx_2=x_2$, $\rho(x_2)\leqq 1$ *and* $\lim_{t\to\infty}(y'_+(t)/|y'_+(t)|)=e_2>0$ *where* $DT(x_2)e_2=\rho(x_2)e_2$.

(E_+) *If* $\lim_{t\to\infty}|y_+(t)|=\infty$ *then* $|T^nx|\to\infty$ *as* $n\to\infty$ *for all* $x\geqq x_1$, $x\neq x_1$.
If $\lim_{t\to\infty}y_+(t)=y_2$ *then* $T^nx\to x_2$ *as* $n\to\infty$ *for all* $x\neq x_1$, $x\in[x_1,x_2]$.

There exists a C^1 function $y_-:[0,\infty)\to\dot{R}^n_-$ satisfying

(A_-) $y_-(t)=x_1-te_1+O(t^2)$ *as* $t\to 0$.

(B_-) $0\leqq t_1<t_2$ *implies* $y_-(t_2)<y_-(t_1)$.

(C_-) $y_-(t)=T(y_-(\rho_1^{-1}t))$, $t\geqq 0$.

(D_-) $\lim_{t\to\infty}y_-(t)=x_0\geqq 0$ *exists,* $Tx_0=x_0$ *and* $\rho(x_0)\leqq 1$. *If $DT(x_0)$ satisfies* (SP) (*e.g.* $x_0>0$) *then* $\lim_{t\to\infty}(y'_-(t)/|y'_-(t)|)=e_0>0$ *where* $DT(x_0)e_0=\rho(x_0)e_0$.

(E_-) $x\neq x_1$, $x\in[x_0,x_1]$ *implies* $T^nx\to x_0$ *as* $n\to\infty$.

Essentially, Theorem 2.5 asserts that a positive unstable fixed point x_1 is joined by two smooth monotone curves, each invariant under T, to two semi-stable fixed points x_0 and x_2 where $0\leqq x_0<x_1<x_2\leqq\infty$. Moreover, the basin of attraction of x_0 includes $[x_0,x_1]$ except for x_1 and the basin of attraction of x_2 includes $[x_1,x_2]$ except for x_1. The two fixed points x_0 and x_2 are asymptotically stable if they are hyperbolic. Observe that if all orbits $O^+(x)$ are bounded then $x_2=\lim_{t\to\infty}y_+(t)$ is finite.

The functional equation which $y_+(t)$ satisfies can be interpreted as follows. Let $C^+=\{y_+(t):t\geqq 0\}$ be the invariant curve joining x_1 to x_2 (or infinity). Then (C_+) asserts that the following diagram of mappings commutes

$$\begin{array}{ccc} C^+ & \xrightarrow{\quad T\quad} & C^+ \\ {\scriptstyle y^+}\uparrow & & \uparrow{\scriptstyle y^+} \\ [0,\infty) & \xrightarrow[t\mapsto\rho_1 t]{} & [0,\infty) \end{array}.$$

In other words, the parametrization y_+ of C^+ affects a linearization of T on C^+. We write C^- for the curve parametrized by y_-. Similar comments apply to C^-.

Proof. As mentioned above, most of the assertions of Theorem 2.5 are contained in [13, Thm. 2.2]. We consider here only those assertions which are not contained in [13]. Of the assertions concerning $y_+(t)$, the strict inequality $y_+(t_1)<y_+(t_2)$ in (B_+) requires proof. In [13] we proved $y_+(t_1)\leqq y_+(t_2)$. The stronger inequality certainly holds for small t by virtue of (A_+). But then, by (C_+), $y_+(t_2)-y_+(t_1)=T^n(y_+(\rho_1^{-n}t_2))-T^n(y_+(\rho_1^{-n}t_1))$ for $n=1,2,\cdots$ and since T^n preserves the strong inequality and $y_+(\rho_1^{-n}t_2)>y_+(\rho_1^{-n}t_1)$ for large n, we deduce that $y_+(t_2)>y_+(t_1)$. The assertions concerning $y_-(t)$ follow for the most part from [13] and the above (apply [13, Thm. 2.2] to $F(x)=x_1-T(x_1-x)$). One assertion which does not follow from [13] is that $y_-(t)$ exists satisfying (A_-), (B_-), (C_-) for all $t>0$, that is, the assertion that $t_0=+\infty$ of [13, Thm. 1.1 (iv)] needs to be verified. Assume $t_0<\infty$, i.e., assume y_- is defined on $[0,t_0)$ satisfying (A_-), (B_-) and (C_-) with $y_-(t)\in\dot{R}^n_+$ holding for $0\leqq t<\rho_1^{-1}t_0$. Since

t_0 is maximal with this property [13, Thm. 1.1(iv)], we have $y_-(\rho_1^{-1}t_0)\in\partial R_+^n$, but then since $y_-(\rho_1^{-1}t_0)=T(y_-(\rho_1^{-2}t_0))$ by continuity of T and $y_-(\rho_1^{-2}t_0)>0$ we have a contradiction to the invariance of $\dot{R}_+^n$ under T. It follows that $t_0=\infty$ and that $0<y_-(t)<x_1$ for $t>0$. Since $y_-(t)$ is bounded below $\lim_{t\to\infty}y_-(t)=x_0\geqq 0$ exists. $Tx_0=x_0$ by continuity of T. (E_-) follows from $\lim_{t\to\infty}y_-(t)=x_0$ as in [13]. If $x_0>0$ then $\rho(x_0)\leqq 1$ for otherwise Theorem 2.5 applied to x_0 yields a contradiction to (E_-). If $x_0\in\partial R_+^n$ one must work harder to show $\rho(x_0)\leqq 1$. In this case we know that $DT(x_0)$ is nonsingular and $DT(x_0)\geqq 0$. By a simple argument involving the normal form of a reducible matrix, we can extend [14, Thm. 2.7] to conclude that there exists a nonzero vector $v\geqq 0$ such that $DT(x_0)v=\rho(x_0)v$ and v does not lie in the range of $DT(x_0)-\rho(x_0)I$. Now one can apply [13, Thm. 1.1] to conclude that if $\rho(x_0)>1$ there exists a monotone curve $z:[0,\infty)\to[x_0,x_1]$, $z(t)=x_0+tv+O(t^2)$ as $t\to 0$, $z(t_1)\leqq z(t_2)$ if $t_1\leqq t_2$ and satisfying (ii), (iii) and $z(t)\to x_1$ as $t\to\infty$. Note (iii) implies that $t_0=\infty$, $z(t)\in[x_0,x_1]$ and $z(\cdot)$ is monotone. Again we have contradicted (E_-).

The positioning of the stable manifold of x_1, $W^s(x_1)=\{x\in\dot{R}_+^n: T^n(x)\to x_1\}$, if nontrivial, was considered by Selgrade [11] for cooperative autonomous systems. Below we state the analogue of this result for mappings which follows immediately from Theorem 2.5.

PROPOSITION 2.6. *Assume x_1 is a hyperbolic, positive fixed point of T with $\rho(x_1)>1$ but assume part of the spectrum of $DT(x_1)$ lies inside the unit circle in the complex plane. Then*

$$W^s(x_1)\cap\left[(x_1+R_+^n)\cup(x_1-R_+^n)\right]=\{x_1\}$$

i.e., no point of $W^s(x_1)$ is weakly related to x_1 other than x_1 itself.

Note that if x is a hyperbolic positive fixed point of the Poincaré map for a cooperative system (1.1) which satisfies the nonpositive divergence condition (1.4) and if $\rho(x)>1$ then it follows immediately that $W^u(x)$ is nontrivial in case $n>1$ since $\det DT(x)\leqq 1$.

If in Theorem 2.5 we make the generic assumption that x_0 and x_2 (if it exists) are hyperbolic fixed points of T, then it follows that they are asymptotically stable. Even in the case that, e.g., $\rho(x_2)=1$ one can obtain information about the stability of x_2. The reason for this is that x_2 possesses a one-dimensional monotone, center manifold which is tangent at x_2 to a positive vector. Our results are contained in the following.

THEOREM 2.7. *Let x_2 be a positive fixed point of T and suppose $\rho(x_2)=1$. Then exactly one of the following holds*:

(i) *T has fixed points $x>x_2$ arbitrarily near x_2.*

(ii) *$B^+(x_2)$ contains the intersection of a ball centered at x_2 with $x_2+R_+^n$.*

(iii) *There is a C^1 curve $C^+\subset x_2+\dot{R}_+^n$ emanating from x_2 which coincides near x_2 with a portion of a center manifold of T at x_2. C^+ is a strictly increasing curve ($x\neq y$, x, $y\in C^+$ then either $x<y$ or $y<x$) which is invariant under T and $Tx>x$ for every $x\in C^+$, $x\neq x_2$. Either C^+ is unbounded in which case $|T^nx|\to\infty$ as $n\to\infty$ for every $x\geqq x_2$ distinct from x_2 or C^+ is bounded and there exists a unique point x_3 satisfying $x_3=\lim_{n\to\infty}T^nx$, independent of $x\in C^+$ distinct from x_2. Moreover x_3 is a fixed point of T, $\rho(x_3)\leqq 1$, and $x_3=\lim_{n\to\infty}T^nx$ for every $x\in[x_2,x_3]$ distinct from x_2.*

Some remarks concerning Theorem 2.7 are in order. In order that its statement not be prohibitively long, Theorem 2.7 describes the situation for $x_2+R_+^n$. An analogous set of alternatives holds for $x_2-R_+^n$ (note x_2 is not necessarily the x_2 of Theorem 2.5). The main difference is that for the analogous assertion corresponding to (iii), $C^-\subset x_2-\dot{R}_+^n$ cannot be unbounded and hence must connect x_2 to a fixed point $x_1<x_2$.

In case x_2 arises from Theorem 2.5 (D_+) and if (ii) of Theorem 2.7 holds, then it follows from the center manifold theorem [1], [9] that x_2 is asymptotically stable, that is, $B(x_2)$ contains a neighborhood of x_2. A similar statement can be made concerning x_0 of Theorem 2.5.

Of course, a one-dimensional center manifold for T containing x_2 exists regardless of which of (i)–(iii) hold. It is most useful in case (iii). It is well known [9] that, in general, there is not a unique center manifold for T corresponding to the fixed point x_2. Our proof of (iii) actually shows that each local center manifold can be extended for $x \geqq x_2$ to a curve C^+. Thus C^+ is not unique. However, Theorem 2.7 implies that if one C^+ is unbounded, then all are unbounded and if one C^+ joins x_2 to x_3 then all such C^+ join x_2 to x_3.

Proof. Suppose (i) does not hold. By (P) and Theorem I there is a positive vector e such that $DT(x_2)e = e$. Since one is a simple eigenvalue and there are no other eigenvalues of $DT(x_2)$ on the unit circle in the complex plane, it follows from the center manifold theorem [1], [9] that there is a C^1, one-dimensional center manifold, tangent at x_2 to e which is locally invariant under T. It follows that we may find a C^1 parametrization (e.g. by arclength) $\hat{x}:[0,\varepsilon) \to x_1 + \dot{R}^n_+$, $\hat{x}(0) = x_2$, $\hat{x}'(0) = e$, of a portion of a center manifold which lies in $x_2 + \dot{R}^n_+$. We assume ε has been chosen so small that $x'(t) > 0$ on $[0,\varepsilon)$. Since T satisfies (M) and the center manifold is locally invariant for T, there exists $\varepsilon_0 \leqq \varepsilon$ sufficiently small so that for each $t \in (0,\varepsilon_0)$ there exists a unique $s \equiv h(t) \in (0,\varepsilon)$ such that $T(x(t)) = x(s)$. One easily sees that $h:[0,\varepsilon_0) \to [0,\varepsilon)$ is strictly increasing and C^1. Our assumption that (i) does not hold implies that either $h(t) < t$ for $0 < t < \varepsilon_0$ or $h(t) > t$ for $0 < t < \varepsilon_0$. In the first case, it is clear from the monotonicity of T that $T^n(x(t)) \to x_2$ as $n \to \infty$ monotonely for every fixed $t \in [0,\varepsilon_0)$. It follows that $B^+(x_2)$ contains $x(t)$ for $0 \leqq t < \varepsilon_0$. It must also contain $[x_2, x]$ for each $x \in B^+(x_2)$ by a familiar argument. Since $x(t) > x_2$ for $0 < t < \varepsilon_0$, we see that the first case, namely that $h(t) < t$, $0 < t < \varepsilon_0$ implies that (ii) holds. The last assertion of (ii) follows from Proposition 2.3.

We now show that the second possibility, $h(t) > t$, $0 < t < \varepsilon_0$ implies that (iii) holds. Fix $t_0 \in (0,\varepsilon_0)$ and for each $n = 0,1,2,\cdots$, define $\hat{x}_n:[0,t_0] \to x_2 + R^n_+$ by $\hat{x}_n(t) = T^n(\hat{x}(t))$, $C_n = \{x : x = \hat{x}_n(t),\ 0 \leqq t \leqq t_0\}$ and $C^+ = \bigcup_{n \geqq 0} C_n$. Observe that $\hat{x}_n$ is C^1 and increasing for each n since $\hat{x}$ is C^1 and increasing and T is C^2 and satisfies (M). We show that $C_0 \subsetneqq C_1 \subsetneqq C_2 \subsetneqq \cdots \subsetneqq C_n \cdots$. Since $C_n = T^n(C_0)$ it suffices to show that $C_0 \subsetneqq C_1$. But $C_0 = \{x : x = \hat{x}(t),\ 0 \leqq t \leqq t_0\}$ and $C_1 = \{x : x = T(\hat{x}(t)),\ 0 \leqq t \leqq t_0\} = \{x : x = \hat{x}(h(t)),\ 0 \leqq t \leqq t_0\} = \{x : x = \hat{x}(t),\ 0 \leqq t \leqq h(t_0)\}$ where $h(t_0) > t_0$. It follows that C^+ is a C^1 curve. Let $x \neq x_2$, $x \in C^+$. Then $x = T^n(\hat{x}(t))$ for some $n \geqq 0$ and $t \in [0,t_0]$ so $Tx = T^{n+1}(\hat{x}(t)) = T^n\hat{x}(h(t)) > T^n(\hat{x}(t)) = x$ since h, $\hat{x}$ and T are increasing. Let $y \neq x_2$, $y \in C^+$, be distinct from x. Then there exists an integer m and distinct values of t, t_x and t_y such that $x = T^m(\hat{x}(t_x))$ and $y = T^m(\hat{x}(t_y))$. Since t_x and t_y are related, it follows that x and y are related. If C^+ is unbounded, then it follows that $|T^n x| \to \infty$ as $n \to \infty$ for $x \in C^+$ distinct from x_2. We want to show that this also holds for $x \geqq x_2$ distinct from x_2. Let $x \geqq x_2$ be distinct from x_2. By Lemma 2.2 there is a positive integer n such that $T^n x > x_2$. It then follows that there exists $y \in C^+$, $y \neq x_2$, such that $T^n x > y$ and hence $T^{n+p}x > T^p y$. The desired conclusion follows from $|T^p y| \to \infty$ as $p \to \infty$. If, on the other hand, C^+ is bounded and $x \in C^+$, $x \neq x_2$, then $x < Tx < T^2x < \cdots$ so $\lim_{n\to\infty} T^n x$ exists. Let $y \in C^+$ be distinct from x_2 and x. If $y > T^n x$ for every n then $y \geqq \lim_{n\to\infty} T^n x$. Since $\lim_{n\to\infty} T^n x$ is a fixed point of T it follows that C^+ contains a fixed point distinct from x_2 which contradicts that $Tx > x$ for $x \in C^+$, $x \neq x_2$. Hence it must be the case that $y < T^n x$ for some integer n and so $\lim_{n\to\infty} T^n y \leqq \lim_{n\to\infty} T^n x$. The symmetry in the argument above implies $\lim T^n x \leqq \lim T^n y$ and so

the limit, x_3, is independent of $x \in C^+$, $x \neq x_2$. By continuity of T, x_3 is a fixed point. Let $x \in [x_2, x_3]$, $x \neq x_2$. By Lemma 2.2 and monotonicity of T it follows that $x_2 < T^n x \leqq x_3$ for large n. But then, for such an n, $T^n x > y$ for some $y \in C^+$ so $T^p y < T^{n+p} \leqq x_3$. Since $T^p y \to x_3$ it follows that $T^n x \to x_3$.

Consider how the results of this section can be combined to give a geometric description of the set of fixed points of a cooperative map T, their domains of attraction and their associated invariant manifolds. We briefly sketch such a description in the case that all fixed points are hyperbolic, there are no unbounded orbits and x_m, the minimal fixed point, is positive. If $x_m > 0$, it follows from Theorem 2.5 that $\rho(x_m) < 1$ so x_m is asymptotically stable. If $B^+(x_m) = x_m + R^n_+$ then $B(x_m) = R^n_+$ by Proposition 2.3, x_m is a global attractor. If $B^+(x_m) \neq x_m + R^n_+$, one expects that the boundary $S^+(x_m)$ will contain fixed points of T. If $B^+(x_m)$ is bounded then by Proposition 2.3, $S^+(x_m)$ will contain at least one fixed point. Such a fixed point x_1 must be unstable with $\rho(x_1) > 1$. In case $\rho(x_1)$ is the only point of the spectrum of $DT(x)$ exceeding one in modulus, then the $(n-1)$-dimensional stable manifold $W^s(x_1)$ must form part of $S^+(x_m)$. Every fixed point of T on $S^+(x_m)$ must be connected to x_m by an invariant monotone curve C^- emanating from the fixed point and belonging to $B^+(x_m)$. The curve C^- must be tangent at x_m to the positive eigenvector for $DT(x_m)$ corresponding to the simple eigenvalue $\rho(x_m)$. If there is more than one fixed point on $S^+(x_m)$, each of the curves C^- are thus tangent at x_m. Each fixed point x_1 of T on $S^+(x_m)$ is connected to another asymptotically stable fixed point $x_2 > x_1$ by an invariant monotone curve C^+ emanating from x_1 and contained in $B^-(x_2)$ by Theorem 2.5. Moreover, $[x_m, x_1] \backslash \{x_1\} \subset B^+(x_m)$ for every such fixed point $x_1 \in S^+(x_m)$ and $[x_1, x_2] \backslash \{x_1\} \subset B^-(x_2)$. It may happen that more than one fixed point on $S^+(x_m)$ is connected by a C^+ to the same fixed point x_2. In that case, each C^+ connecting a fixed point on $S^+(x_m)$ to x_2 is tangent at x_2 to the positive eigenvector for $DT(x_2)$.

Now each of these secondary stable fixed points x_2 can take the place of x_m in the previous scenario. Thus one obtains a cascading sequence of these stable-unstable-stable cells forming a tree-like structure with x_m at its base. In §4 we will obtain additional information for two-dimensional cooperative maps which will allow us to describe completely their "phase portrait".

3. Fixed points of competitive maps. In this brief section we examine how the fixed points of competitive maps and their invariant manifolds are situated by applying the results of the previous section to the inverse map. A map T will be called a competitive map if T is a C^2 diffeomorphism defined on a neighborhood of R^n_+ satisfying $Tx > 0$ if and only if $x > 0$ and

(MI) $x_1, x_2 \geqq 0$ and $Tx_1 < Tx_2$ imply $x_1 < x_2$

and

(SPI) $x > 0$ and $T(x) = x$ implies $DT(x)^{-p} > 0$ for some positive integer p depending on x.

(MI) and (SPI) will be assumed without further mention but we will briefly note a result requiring the stronger monotonicity assumption

(SMI) $x_1 \geqq 0$, $x_2 > 0$ and $Tx_1 \leqq Tx_2$ implies $x_1 < x_2$.

If T is the Poincaré map of a C^2 competitive system (1.1) then (MI) holds. If, in addition, (1.1), is irreducible then (SPI) and (SMI) hold for T by Corollary F and Lemma H respectively. Of course, (MI), (SPI) and (SMI) imply that T^{-1}, the inverse of T, satisfies (M), (SP) and (SM) of the previous section except that T^{-1} is defined only in a neighborhood of $T(R^n_+)$. In case T is the Poincaré map of a competitive system (1.1) then Corollary G implies that $[0, x] \subset T(R^n_+)$ whenever $x \in T(R^n_+)$.

We note that competitive maps necessarily fix the origin as might be expected from the remark following Corollary C.

PROPOSITION 3.1. *Let T be a competitive map. Then $T0=0$. If x_1 and x_2 are distinct fixed points of T, at least one of which is positive, then $x_1 \leqq x_2$ implies $x_1 < x_2$.*

Proof. $T(\partial R^n_+) \subset \partial R^n_+$ together with the smoothness of T and the fact that $DT(0)$ is nonsingular imply $T0=0$. The second assertion follows from Lemma 2.2 applied to T^{-1}.

As previously mentioned, the results of the previous section concerning cooperative maps can be immediately applied to T^{-1} to obtain results for competitive maps. It is not our intention to translate each of the results of the previous section to a corresponding result for competitive maps. Instead, we select a few such results which are likely to be of interest for competitive Poincaré maps. Keep in mind as we proceed that if T is a competitive map and $Tx=x$ then T^{-1} is defined in a neighborhood of x, $T^{-1}x=x$ and the spectrum of $D(T^{-1})(x)$ is obtained from the spectrum of $DT(x)$ by inverting: $\lambda \to \lambda^{-1}$. If x is a positive fixed point of T, let $\mu(x)=[\rho(D(T^{-1}))(x)]^{-1}$. By (SPI) and Theorem I, $\mu(x)^{-1}$ is a simple eigenvalue of $D(T^{-1})(x)=(DT(x))^{-1}$ with positive eigenvector and all other eigenvalues have smaller modulus. It follows that $\mu(x)$ is the (strictly) smallest eigenvalue of $DT(x)$ in modulus and it has a positive corresponding eigenvector.

If x is a positive fixed point of T with $\mu(x)>1$ then, of course, x is a totally unstable fixed point of T and it is an asymptotically stable fixed point of T^{-1}. Consequently, Proposition 2.3 (and Proposition 2.4 if (SMI) holds) can be applied to obtain information concerning the domain of repulsion for x and its boundary. It is much more likely, however, that $\mu(x)<1$. In fact, if a competitive system (1.1) has a nonpositive divergence ((1.4) holds) then every positive fixed point of its Poincaré map satisfies $\mu<1$.

LEMMA 3.2. *If $n>1$ and $x>0$ is a fixed point of T and $\det(DT(x)) \leqq 1$ then $\mu(x)<1$.*

Proof. If $\mu(x) \geqq 1$ then since $\mu(x)$ is the (strictly) smallest eigenvalue it follows that the products of the eigenvalues of $DT(x)$ with corresponding multiplicities strictly exceeds ($n>1$) one which contradicts $\det(DT(x)) \leqq 1$.

If $\mu(x)<1$ for a positive fixed point x of T then the spectral radius of $D(T^{-1})(x)$ exceeds one and Theorem 2.5 can be applied to T^{-1} to obtain the following result.

THEOREM 3.3. *Let x_1 be a positive fixed point of T with $\mu_1=\mu_1(x_1)<1$ and $DT(x_1)e_1=\mu_1 e_1$ where $e_1>0$. Then there exists t_0, $0<t_0 \leqq \infty$, and a C^1 function $y_+ : [0,t_0) \to \dot{R}^n_+$ satisfying*

(A_+) *$y_+(t)=x_1+te_1+O(t^2)$ as $t \to 0$.*

(B_+) *$0<t_1<t_2<t_0$ implies $y_+(t_1)<y_+(t_2)$.*

(C_+) *$T(y_+(t))=y_+(\mu_1 t)$, $0<t<t_0$.*

(D_+) *Either $\lim_{t \to t_0+}|y_+(t)|=\infty$ or $\lim_{t \to t_0+} y_+(t)=x_2$. In the latter case, $t_0=\infty$, $Tx_2=x_2$, $\mu(x_2) \geqq 1$ and $\lim_{t \to \infty}(y'_+(t)/|y'_+(t)|)=e_2>0$ where $DT(y_2)e_2=\mu(x_2)e_2$.*

(E_+) *If $\lim_{t \to t_0+}|y_+(t)|=\infty$ then for all $x \geqq x_1$, $x \neq x_1$, either there exists N such that $T^{-\bar{n}}x \in T(R^n_+)$, $0 \leqq \bar{n} \leqq N$ and $T^{-(\bar{n}+1)}x \notin T(R^n_+)$ or $T^{-\bar{n}}x \in T(R^n_+)$ for all $\bar{n}$ and $|T^{-\bar{n}}x| \to \infty$ as $\bar{n} \to \infty$. If $\lim_{t \to \infty} y_+(t)=x_2$ then $T^{-\bar{n}}x \to x_2$ as $\bar{n} \to \infty$ for all $x \neq x_1$, $x \in [x_1,x_2]$.*

There exists a C^1 function $y_- : [0,\infty) \to \dot{R}^n_+$ satisfying

(A_-) *$y_-(t)=x_1-te_1+O(t^2)$ as $t \to 0$.*

(B_-) *$0 \leqq t_1<t_2$ implies $y_-(t_2)<y_-(t_1)$.*

(C_) $T(y_-(t))=y_-(\mu_1 t)$, $t\geqq 0$.

(D_) $\lim_{t\to\infty} y_-(t)=x_0\geqq 0$ *exists,* $Tx_0=x_0$ *and* $\mu(x_0)\geqq 1$. *If* $DT(x_0)$ *satisfies* (SPI) (*e.g.* $x_0>0$) *then* $\lim_{t\to\infty}-(y'_-(t)/|y'_-(t)|)=e_0>0$ *where* $DT(x_0)e_0=\mu(x_0)e_0$.

(E_) $x\neq x_1$, $x\in[x_0,x_1]$ *implies* $T^{-n}x\to x_0$ *as* $n\to\infty$.

As an immediate corollary of Theorem 3.3 (E_) observe that no positive asymptotically stable fixed point x_1 of T can exist unless T possesses a corresponding fixed point $x_0\leqq x_1$ which satisfies $\mu(x_0)\geqq 1$. Ignoring the nongeneric possibility $\mu(x_0)=1$, one may say that x_0 is a repeller: all of the eigenvalues of the Jacobian at x_0 exceed unity in modulus.

In view of the fact that $\mu(x)<1$ is typical for a positive fixed point (and necessary if $\det DT(x)\leqq 1$), Theorem 3.3 will play a more fundamental role for competitive maps than its counterpart, Theorem 2.5, played for cooperative maps. The assertions (A_)–(E_) will be most useful in general. Let us write $C^+=\{y_+(t):0\leqq t<t_0\}$ and $C^-=\{y_-(t):t\geqq 0\}$ for the two invariant curves the existence of which is asserted in Theorem 3.3. We call special attention to the fact that assertions made in (E_+) and (E_) concern T^{-1} and not T itself. For example, in (E_) the set $[x_0,x_1]$ is mapped into itself by T^{-1} and the orbit, under T^{-1}, of every point except x_1 of $[x_0,x_1]$ limits on x_0 but $[x_0,x_1]$ is not necessarily mapped into itself by T. It is also important to reemphasize that when either of the curves C^+ or C^- connect x_1 to a fixed point x_2 or x_0, then that fixed point must be quite unstable since the smallest eigenvalue of the Jacobian must be greater than or equal to one. One might expect that for a typical application there are few such ($\mu\geqq 1$) unstable fixed points. Indeed, suppose that T is the Poincaré map for a competitive irreducible system satisfying the nonpositive divergence condition (1.4). Then by Lemma 3.2, every positive fixed point x of T has $\mu(x)<1$. This observation excludes the possibility that $\lim_{t\to\infty} y_+(t)=x_2$ in (D_+) and that $x_0>0$ in (D_) of Theorem 3.3. One can, in fact, say slightly more.

PROPOSITION 3.4. *Let* T *be a competitive map with the property that* $\mu(x)<1$ *for each of its positive fixed points. Then no pair of positive fixed points of* T *can be weakly related and every positive fixed point* x_1 *is joined by the curve* C^- *to a fixed point* x_0 *of* T *on* ∂R^n_+. *Moreover,* $\mu(x_0)\geqq 1$ *and* $T^{-\bar n}x\to x_0$ *as* $\bar n\to\infty$ *for every* $x\in[x_0,x_1]$ *distinct from* x_1.

Proof. If $0<x_0\leqq x_2$ where x_0 and x_2 are fixed points of T, then $x_0<x_2$ by Proposition 3.1 and Theorem 3.2 (D_+) applied to x_0 implies the existence of a fixed point $x_1\in[x_0,x_2]$ with $\mu(x_1)\geqq 1$. Indeed, the curve C^+ obtained from Theorem 3.2 applied to x_0 must satisfy $C^+\subset[x_0,x_2]$ by Remark 3 following Theorem 1.1 in [13]. The first assertion is established. The second assertion is a restatement of Theorem 3.2 (E_).

Anyone who has worked with competitive systems (1.1) in applications knows that typically each face $H_I^+=\{x\geqq 0:x_i=0,\ i\in I\}$, $I\subseteq\{1,2,\cdots,n\}$, making up the boundary of R^n_+ is invariant under (1.1). In such a case, the Poincaré map will map each H_I^+ into itself. We will call a competitive map T "competitive in each face" if T is a competitive map when restricted to each face H_I^+ (where $\leqq$ and $<$ are now relative to H_I). Of course, each of the results of this section may be applied to each of the faces H_I^+ if T is competitive in each face. For example, if T is competitive in each face, then no face H_I^+ can have a fixed point x_1 belonging to the interior of H_I^+ (relative to H_I) which is asymptotically stable for $T|_{H_I^+}:H_I^+\to H_I^+$ unless $T|_{H_I^+}$ possesses a fixed point $x_0\leqq x_1$ in H_i^+ which satisfies $\mu_{H_I^+}(x_0)\geqq 1$ where $\mu_{H_I^+}(x_0)$ is the smallest eigenvalue of $D(T|_{H_I^+})(x_0)=DT(x_0)|_{H_I^+}:H_I\to H_I$.

If T is competitive in each face, then each of the coordinate axes is invariant under T. It follows that $DT(0)$ is a diagonal matrix. In many applications to ecology where competitive systems arise, each coordinate of x represents the population density of a particular organism and typically each organism is assumed to be viable in the sense that if none of its competitors are present, then it will grow, at least when it is scarce. The mathematical translation of this last statement is that if $DT(0) = \operatorname{diag}\{\lambda_1, \lambda_2, \cdots, \lambda_n\}$ then $\lambda_i \geqq 1$, $1 \leqq i \leqq n$. Typically, one expects 0 to be the only fixed point of T for which $\mu \geqq 1$ and thus all fixed points of T will be connected to 0 by a curve C^-.

An immediate corollary of Theorem 3.3 is the following.

PROPOSITION 3.5. *Assume x_0 is a hyperbolic, positive fixed point of a competitive map T with $\mu(x_0) < 1$ and $\rho(DT(x_0)) > 1$. Then no point of the unstable manifold $W^u(x_0) = \{x : T^{-\bar{n}}x \to x_0 \text{ as } \bar{n} \to \infty\}$ is weakly related to x_0 except x_0 itself.*

In future work we intend to consider in greater detail the dynamics generated by maps which are competitive in each face in low dimensions ($n = 2$ and 3).

4. Cooperative planar maps. In this section we continue our investigation of cooperative maps, focusing on the two-dimensional case. The results obtained in this section may be applied to the inverse of a competitive planar map.

In addition to the results of §2, we will show that the dynamics of a planar orientation preserving cooperative map are "trivial" in the sense that all bounded orbits tend to fixed points. This result was obtained by de Mottoni and Schiaffino [10] for competitive maps and we essentially reproduce their proof with minor modifications. We also show that the center-stable manifold of every fixed point contains two monotone invariant curves each of which either tends to infinity or limits on another fixed point of T. Armed with these additional results together with those of §2, we will be able to describe completely the "phase portrait" for a planar orientation preserving cooperative map which satisfies (A) the minimal fixed point x_m is positive, (B) all fixed points are hyperbolic and (C) all orbits are bounded. A rather rich structure will become apparent for the set of fixed points and their basins of attraction. These results are contained in Theorem 4.8, the main result of this section.

Some additional notation will prove useful. Let Q_i, $i = 1, 2, 3, 4$ denote the usual open quadrants in R^2 in counter clockwise order with increasing i, e.g., $Q_1 = \{(x_1, x_2) : x_i > 0,\ i = 1, 2\}$. For $x \geqq 0$ denote by $Q_i(x)$ the set $(x + Q_i) \cap R^2_+$, that is, the portion of the ith quadrant centered at x which lies in R^2_+.

The manner in which orientation enters the study of cooperative maps may be seen in the following result. First, observe that if T is cooperative then $T(Q_i(x)) \subset Q_i(Tx)$ for $i = 1, 3$.

LEMMA 4.1. *Let T be cooperative and $x > 0$. If T is orientation preserving, then $T(Q_2(x)) \cap Q_4(Tx)$ and $T(Q_4(x)) \cap Q_2(Tx)$ are empty. If T is orientation reversing, then $T(Q_2(x)) \cap Q_2(Tx)$ and $T(Q_4(x)) \cap Q_4(Tx)$ are empty.*

Proof. The result is easily established by considering the images under T of the following curves. $E_1 = \{(t, 0) : t \geqq 0\}$, $E_2 = \{(0, t) : t \geqq 0\}$, $H_1 = \{(t, b) : t \geqq 0\}$ and $H_2 = \{(a, t) : t \geqq 0\}$ where $x = (a, b)$. Each of the curves has been parametrized in such a way as to be monotone nondecreasing. Each of the image curves $T(E_1)$, $T(E_2)$, $T(H_1)$ and $T(H_2)$ are monotone nondecreasing and $T(H_i)$, $i = 1, 2$ lie in $\overline{Q_1(Tx)} \cup \overline{Q_3(Tx)}$. The proposition follows easily from the fact that the boundary of $T(Q_2(x))$ consists of a portion of each of the curves $T(H_1)$, $T(H_2)$ and $T(E_2)$ and the fact that the relative relation of the pairs of curves $T(E_1)$, $T(E_2)$ and $T(H_1)$ and $T(H_2)$ are preserved or reserved depending on whether T is orientation preserving or reversing.

PROPOSITION 4.2. *If T is a cooperative orientation preserving map and $x>0$ then $T(Q_i(x))\cap Q_i(Tx)$ $i=1,3$ and $T(Q_j(x))\cap Q_k(Tx)=\phi$, $j\neq k$, $j,k\in\{2,4\}$.*

Proposition 4.2 has the important implication that the orbit of x, $0^+(x)$ has eventually monotone (with respect to n) components. This line of argument was first used by de Mottoni and Schiaffino in the competitive case and our proof below is merely a mirror image of theirs.

THEOREM 4.3. *Let T be cooperative and orientation preserving, $x=(x_1,x_2)>0$ and $x_n=T^nx=(x_1^{(n)},x_2^{(n)})$, $n=0,1,2,\cdots$. Then there exists $N=N(x)$, a positive integer, such that for $n\geqq N$ both $x_1^{(n)}$ and $x_2^{(n)}$ are monotone sequences.*

Proof. Either there exists a nonnegative integer M such that $T^Mx\leqq T^{N+1}x(T^{N+1}x\leqq T^Nx)$ or it must be the case that $T^{n+1}x\in Q_2(T^nx)\cup Q_4(T^nx)$ for all $n\geqq 0$. In the former case, $T^nx\leqq T^{n+1}x(T^{n+1}x\leqq T^nx)$ for all $n\geqq N$ and the theorem follows. In the latter case we may assume $Tx\in Q_2(x)$ (the case $Tx\in Q_4(x)$ is treated similarly). It follows from Lemma 4.1 that $T^2x\in T(Q_2(x))$ does not lie in $Q_4(Tx)$ and so must be an element of $Q_2(Tx)$. Proceeding by induction on n, one shows $T^{n+1}x\in Q_2(T^nx)$ $n=0,1,2,\cdots$. The result follows easily in this case (in fact $N=0$).

The upshot of Theorem 4.3 is that the dynamics of cooperative orientation preserving maps is "trivial", all bounded orbits are either fixed points or tend to fixed points. Though the dynamics may be "trivial", we will show that the fixed point set of an orientation preserving cooperative map together with their domains of attraction have a rich and beautiful structure.

COROLLARY 4.4. *Let T be cooperative and orientation preserving and let $x>0$. Then either $|T^nx|\to\infty$ as $n\to\infty$ or there exists a fixed point $x_1\geqq 0$ of T for which $T^nx\to x_1$ as $n\to\infty$.*

If T is a cooperative map, it may not be the case that T^2 is cooperative. The problem is that (SP) may not hold for T^2. In case (SP) holds for T^2 so it is cooperative, then T^2 is cooperative and orientation preserving. Hence T^2 has trivial dynamics. In this way, one can treat orientation reversing cooperative maps: their bounded orbits approach period two points (which might be fixed points) of T.

Let $x>0$ be a fixed point of an orientation preserving cooperative map T. Since $DT(x)$ satisfies (SP), the spectrum of $DT(x)$ consists of distinct positive eigenvalues η, ρ where $\rho=\rho(DT(x))$ is the spectral radius and $\eta<\rho$. Corresponding to ρ there is a positive eigenvector e_1. This eigenvector, being (to within scalar multiple) the unique nonnegative eigenvector of $DT(x)$ by Theorem 1, it follows that we may select an eigenvector e_2 corresponding to η which lies in Q_2 or Q_4 as we please. In case $\rho>1$, Theorem 2.5 implies the existence of two monotone invariant curves for T: $C^+\subset Q_1(x)$ and $C^-\subset Q_3(x)$. The following result states that if $\eta<1$ then there are two monotone invariant curves for T: $C^l\subset Q_2(x)$ and $C^r\subset Q_4(x)$.

THEOREM 4.5. *Let x_1 be a positive fixed point of an orientation preserving cooperative map T and suppose $\eta_1=\eta(x_1)$, the smaller positive eigenvalue of $DT(x_1)$ satisfies $\eta_1<1$. Let $DT(x_1)e_2=\eta_1e_2$ where the eigenvector $e_2\in Q_2$. Then there exist t_l, $0<t_l\leqq\infty$ and a C^1 function $y_l\colon[0,t_l)\to Q_2(x_1)$ satisfying*

- (A$_l$) *$y_l(t)=x_1+te_2+O(t^2)$ as $t\to 0$.*
- (B$_l$) *$0<t<s<t_l$ implies $y(s)-y(t)\in Q_2$.*
- (C$_l$) *$T(y_l(t))=y_l(\eta_1t)$, $0\leqq t<t_l$, and either $t_l=\infty$ or $y_l(t_l+)$ lies on the vertical coordinate axis.*
- (D$_l$) *If $t_l=\infty$ then either $\lim_{t\to\infty}y_l(t)=\infty$ or $\lim_{t\to\infty}y_l(t)=x_l\in Q_2(x_1)$. In the latter case, $Tx_l=x_l$ and, if $x_l>0$, then $\eta(x_l)\geqq 1$ and $\lim_{t\to\infty}(y_l'(t)/|y_l'(t)|)$ is an eigenvector for $DT(x_l)$ corresponding to $\eta(x_l)$.*

(E_l) *If* $x \geqq y_l(t)$ *for some* $t \in (0, t_l)$ *then* $T^n x \geqq y_l(\eta_1^n t)$ *for* $n = 1, 2, \cdots$.
If $x \leqq y_l(t)$ *for some* $t \in (0, t_l)$ *then* $T^n x \leqq y_l(n_1^n t)$ *for* $n = 1, 2, \cdots$.

There exist t_r, $0 < t_r \leqq \infty$ *and a* C^1 *function* $y_r : [0, t_r) \to Q_4(x_1)$ *satisfying*

(A_r) $y_r(t) = x_1 - te_2 + O(t^2)$ *as* $t \to 0$.

(B_r) $0 < t < s < t_r$ *implies* $y_r(s) - y_r(t) \in Q_4$.

(C_r) $T(y_r(t)) = y_r(\eta_1 t)$ $0 \leqq t < t_r$, *and either* $t_r = \infty$ *or* $y_r(t_r+)$ *lies on the horizontal coordinate axis.*

(D_r) *If* $t_r = \infty$ *then either* $\lim_{t\to\infty} y_r(t) = \infty$ *or* $\lim_{t\to\infty} y_r(t) = x_r \in Q_4(x_1)$. *In the latter case,* $Tx_r = x_r$ *and, if* $x_r > 0$ *then* $\eta(x_r) \geqq 1$ *and* $\lim_{t\to\infty} (y_r'(t)/|y_r'(t)|)$ *is an eigenvector for* $DT(x_r)$ *corresponding to* $\eta(x_r)$.

(E_r) *If* $x \geqq y_r(t)$ *for some* $t \in (0, t_r)$ *then* $T^n x \geqq y_r(\eta_1^n t)$ *for* $n = 1, 2, \cdots$.
If $x \leqq y_r(t)$ *for some* $t \in (0, t_r)$ *then* $T^n x \geqq y_r(\eta_1^n t)$ *for* $n = 1, 2, \cdots$.

Before proceeding to the proof of Theorem 4.5 we comment on its assertions which clearly parallel those of Theorem 2.5. Figure 4.1 illustrates the possible configurations of the curve $C^l \equiv \{ y_l(t) : 0 \leqq t < t_l \}$ which lies in the center-stable manifold of the fixed point x_1 of T. The curve $C^r \equiv \{ y_r(t) : 0 \leqq t < t_r \}$ has a similar set of possibilities which are independent of those of C^l. In case the curve C^l connects x_1 to a positive fixed point x_l, see Fig. 4.1 (b), (D_l) asserts that the smallest eigenvalue of the Jacobian of T at x_l is larger than or equal to one. In other words, x_l is a repeller if it is hyperbolic.

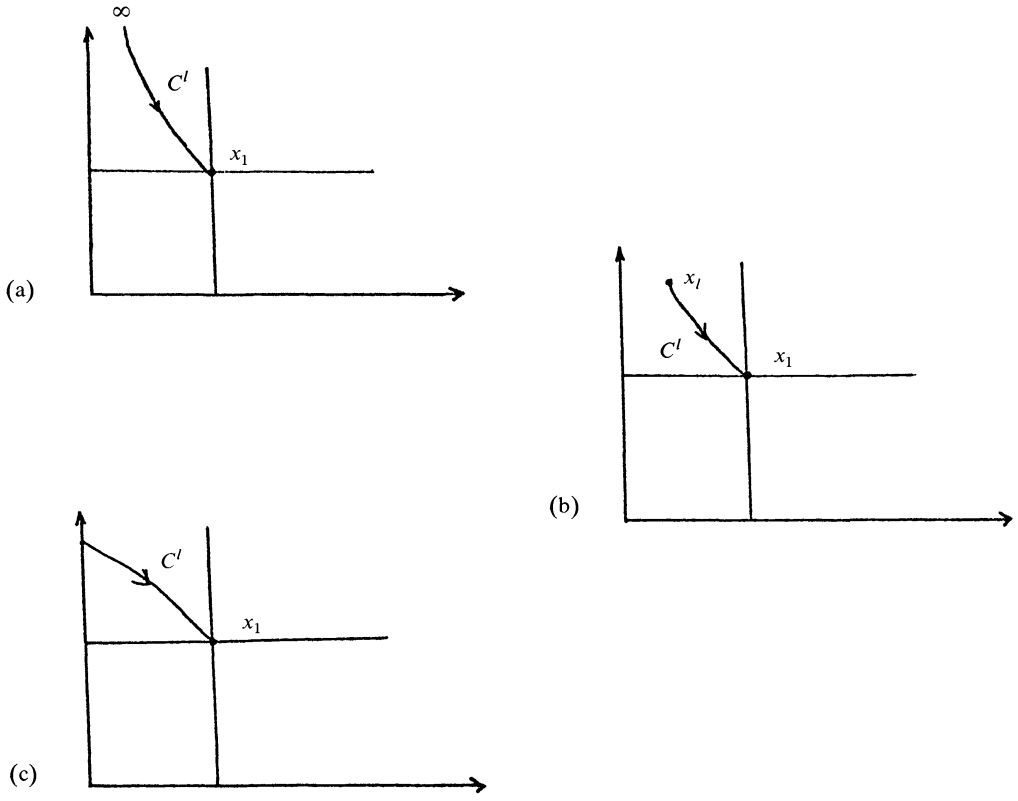


FIG. 4.1. *The possible configurations of* C^l. *In* (a) $t_l = \infty$ *and* $\lim_{t\to\infty} y_l(t) = \infty$; *in* (b) $t_l = \infty$ *and* $\lim_{t\to\infty} y_l(t) = x_l$. (c) *illustrates two possibilities: either* $t_l = \infty$ *and* $\lim_{t\to\infty} y_l(t) = x_l \in \partial R_+^2$ *or* $t_l < \infty$ *and* $y_l(t_l+)$, *which cannot be a fixed point, lies on* ∂R_+^2.

The assertions (E_l) and (E_r) will prove to be especially important. Since $y_l(\eta_1^n t) \to x_1$ as $n \to \infty$, if $x \geqq y_l(t)$ then any limit points of $\{T^n x\}_{n \geqq 0}$ must lie in $\overline{Q_1(x_1)}$. Similar remarks hold for the other cases.

Proof. Let E_1 and E_2 denote the horizontal and vertical coordinate axis respectively. Since T satisfies (M), $T(E_i)$, $i=1,2$, are monotone curves forming the boundary of $T(Q_1)$. See Fig. 4.2 depicting these sets. Consider $T^{-1}: T(Q_1) \to Q_1$ which fixes x_1 and for which $\rho(D(T^{-1})(x_1)) = \eta_1^{-1} > 1$. Applying [13, Thm. 1.1] to T^{-1} with eigenvector $e_2 \in Q_2(x_1)$, we obtain t_l, $0 < t_l \leqq \infty$ and a C^1 function $y_l: [0, t_l) \to T(Q_1)$ satisfying $y_l(t) = x_1 + te_2 + O(t^2)$ as $t \to 0$, $y_l(t) = T^{-1}(y_l(\eta_1 t))$, $0 \leqq t < t_l$, and $t_l = \infty$ or t_l is maximal with the property that $y_l(t) \in T(Q_1)$, $0 \leqq t < \eta_1 t_l$. Since $y_l'(t) \in Q_2$ for sufficiently small $t > 0$, $y_l(t) \in Q_2(x_1)$ and (B_l) hold for small t. Suppose $0 < t < s < t_l$ and $y_l(t)$ and $y_l(s)$ are weakly related, e.g., $y_l(t) \leqq y_l(s)$. Then $T(y_l(t)) \leqq T(y_l(s))$ so $y_l(\eta_1 t) \leqq y_l(\eta_1 s)$. After n applications of T one has $y_l(\eta_1^n t) \leqq y_l(\eta_1^n s)$ which contradicts, for large enough n $(\eta_1^n \ll 1)$, that B_l holds for small t. Now let $F: \{(t,s): 0 < t < s < t_l\} \to R^2 \setminus \{0\}$ be defined by $F(t,s) = y_l(s) - y_l(t)$. F is a continuous function satisfying Image $F \subset Q_2 \cup Q_4$ by the argument just presented. Since B_l holds for small t, Image $F \cap Q_2 \neq \varnothing$. But Image F must be connected and so it must be the case that Image $F \subset Q_2$. Hence (B_l) holds and $y_l(t) \in Q_2(x_1)$, $0 < t < t_l$. Now, if $t_l < \infty$ then (B_l) implies the existence of $y_l(t_l+) \in Q_2(x_1)$ and since $t_l > 0$ is maximal with the property that $y_l(t) \in T(Q_1)$, $0 \leqq t < \eta_1 t_l$, it follows that $y_l(\eta_1 t_l) \in \partial T(Q_1) = T(E_1) \cup T(E_2)$. Since $y_l(\eta_1 t_l) \in Q_2(x_1)$, $y_l(\eta_1 t_l) \in T(L_2)$ (see Fig. 4.1). Hence $y_l(t_l+) \in E_2$. This completes the proof of (C_l). If $t_l = \infty$ then (B_l) implies the first assertion of (D_l). Clearly $Tx_l = x_l$. Before proving the remainder of (D_l) we observe that (E_l) follows trivially from (C_l) and (M). Now suppose $x_l > 0$ and $\eta(x_l) < 1$. But then we could apply Theorem 4.5 $(A_r) \to (E_r)$ to x_l. This contradicts (E_l) (draw a picture!). The last assertion of (D_l) can be verified as in [13, Thm. 2.2]. The proof of $(A_r) \to (E_r)$ follows a similar pattern.

Theorem 2.5, Corollary 4.4 and Theorem 4.5 can be used to describe the "phase portrait" for a planar orientation preserving cooperative map satisfying

A. The minimal fixed point, x_m, is positive.

B. Every fixed point is hyperbolic.

C. All forward orbits, $O^+(x)$, are bounded.

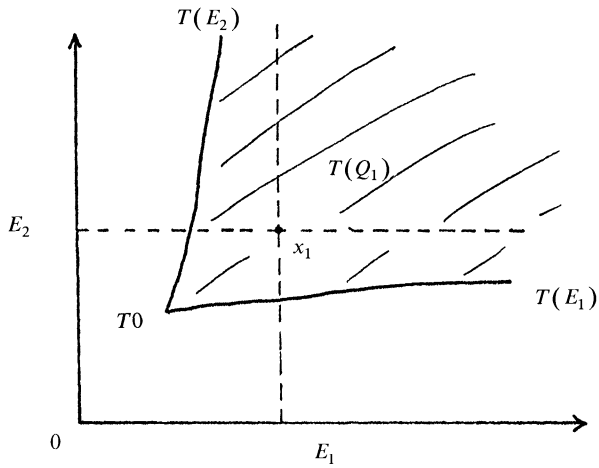


FIG. 4.2. *The configuration of $T(E_i)$, $i=1,2$ and $T(Q_1)$. Note $T0$ need not be positive as depicted. Each curve $T(E_i)$ is monotone.*

In view of B, a fixed point x of T must be either (i) stable: $0<\eta<\rho<1$, (ii) a saddle point: $0<\eta<1<\rho$, or (iii) a repeller: $1<\eta<\rho$. If x is a positive fixed point of T and either stable or a saddle point, we will write $C^l(x)$ and $C^r(x)$ for the two monotone invariant curves emanating from x which are asserted to exist by Theorem 4.5. Similarly, if x is a saddle point or a repeller, we write $C^-(x)$ and $C^+(x)$ for the monotone invariant curves emanating from x described in Theorem 2.5. Observe that, by B and C, a $C^-(x)$ and $C^+(x)$ always lead to a stable fixed point while if a $C^l(x)$ or $C^r(x)$ lead to a fixed point, then that fixed point must be a repeller. We will establish the following propositions assuming that A–C hold.

PROPOSITION 4.6. *If y_1 is a stable positive fixed point then either $B^+(y_1)=\overline{Q_1(y_1)}$ or $S^+(y_1)$ is the graph of a C^1 monotone decreasing curve in $\overline{Q_1(y_1)}$. $S^+(y_1)$ contains an odd number of fixed points $x_1, x_2, \cdots, x_{2n+1}$ ordered from left to right on $S^+(y_1)$ (see Fig. 4.3). The odd indexed points are saddles and the even indexed points are repellers. $S^+(y_1)$ consists of the curves $C^l(x_i)$, $C^r(x_i)$, $i=1,3,5,\cdots,2n-1$ together with their limit points $x_2, x_4, \cdots, x_{2n}$.*

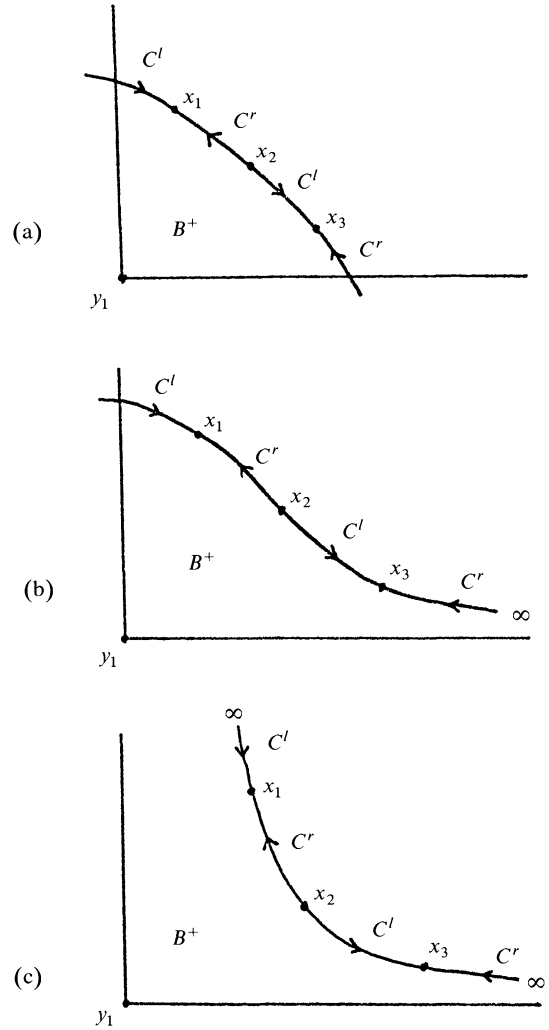


FIG. 4.3. *The set $S^+(y_1)$ in case $2n+1=3$.*

PROPOSITION 4.7. *If y_1 is a stable fixed point, $B^+(y_1) \neq \overline{Q_1(y_1)}$, and $x_1, x_2, \cdots, x_{2n+1}$ are the fixed points of T on $S^+(x)$ as in Proposition* 4.6, *then each of the curves $C^-(x_i)$, $1 \leqq i \leqq 2n+1$, is asymptotic to y_1 and tangent at y_1 to the positive eigenvector for $DT(y_1)$ (corresponding to $\rho(y_1)$). All the curves $C^+(x_i)$, $1 \leqq i \leqq 2n+1$, are asymptotic to the same stable fixed point $y_2 > x_i$, $1 \leqq i \leqq 2n+1$. Each curve $C^+(x_i)$ is tangent at y_2 to the positive eigenvector for $DT(y_2)$. The curve $C^l(y_2) \subset Q_2(y_2)$ either leaves $\overline{Q_1(y_1)}$ by crossing its vertical boundary or remains in $Q_1(y_1)$ tending to infinity. The curve $C^r(y_2) \subset Q_4(y_2)$ either leaves $\overline{Q_1(y_1)}$ by crossing its horizontal boundary or remains in $Q_1(y_1)$ tending to infinity. (See Fig.* 4.4.)

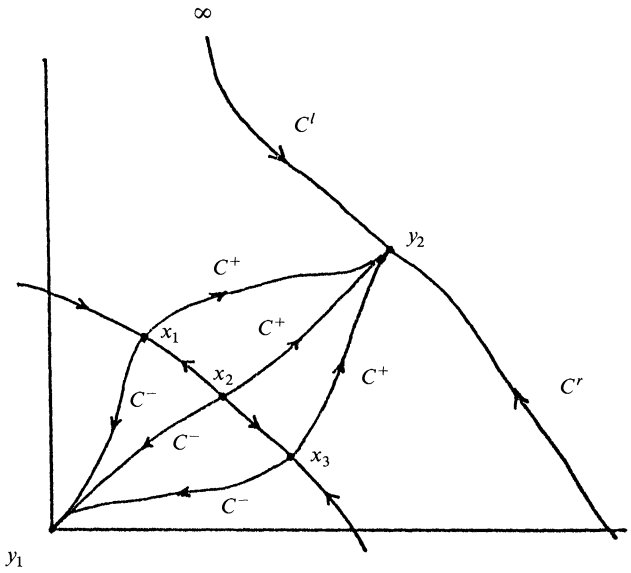


FIG. 4.4. *The curves $C^+(x_i)$, $C^-(x_i)$, $C^r(y_2)$ and $C^l(y_2)$ are added to Fig.* 4.3(a). *The reader may easily supply these curves to Fig.* 4.3(b) *and* (c). *Note that no two distinct pair of the various curves can intersect so if $C^l(x_1)$ is unbounded in $Q_1(y_1)$ then so is $C^l(y_2)$.*

Propositions 4.6 and 4.7 describe what we might refer to as a basic cell $[y_1, y_2]$ with y_1 and y_2 being stable fixed points. Inside the cell there are an odd number of unstable fixed points lying on a smooth monotone decreasing curve (only part of which may lie in $[y_1, y_2]$). The proofs of Propositions 4.6 and 4.7 will be deferred to the end of this section while we consider how to construct the global phase portrait. We will show that the fixed point set of T consists of a finite or infinite totally ordered sequence of such basic units and we will describe precisely the basis of attraction of each of the stable fixed points. In order to achieve this goal, we proceed from the bottom up, so to speak, beginning with the minimal fixed point $y_1 \equiv x_m$. In Fig. 4.5 below, we indicate the possible configurations of the curves $C^l(y_1)$ and $C^r(y_1)$. Observe that since all fixed points x of T satisfy $x \geqq y_1$, the quadrants $\overline{Q_2(y_1)}$ and $\overline{Q_4(y_1)}$ are fixed point free while $\overline{Q_3(y_1)} \subset B^-(y_1)$.

By Propositions 4.6 and 4.7, either $B^+(y_1) = \overline{Q_1(y_1)}$ or we can add one of our basic units as in Fig. 4.4 to Fig. 4.5. In case $B^+(y_1) = \overline{Q_1(y_1)}$, then $B(y_1) = \overline{Q}_1$, all orbits tend to y_1. If $B^+(y_1) \neq \overline{Q_1(y_1)}$ so that Propositions 4.6 and 4.7 apply, there are two issues to consider. First, where do the curves $C^l(x_1)$, $C^l(y_2)$, $C^r(x_{2n+1})$ and $C^r(y_2)$ go if they leave $\overline{Q_1(y_1)}$? Secondly, are there unaccounted-for fixed points of T belonging to the various regions partitioned by the curves which we have described so far?

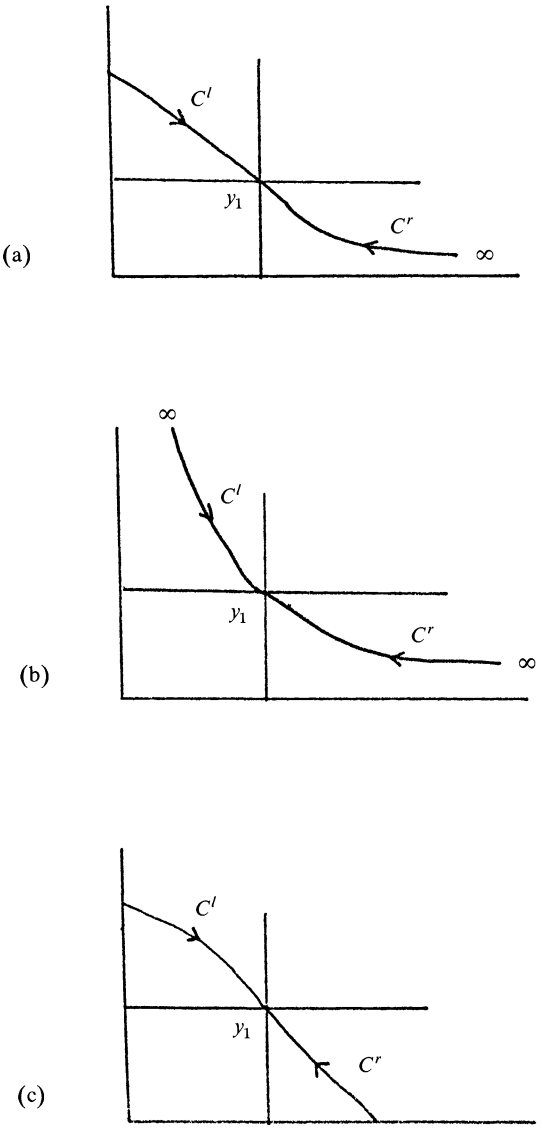


FIG. 4.5. *Configurations of $C^l(y_1)$ and $C^r(y_1)$ for the minimal fixed point y_1. $\overline{Q_2(y_1)}$, $\overline{Q_3(y_1)}$ and $\overline{Q_4(y_1)}$ are fixed point free.*

Beginning with the first question, recall that $C^l(x_1)$ either leaves $\overline{Q_1(y_1)}$ through the vertical boundary or remains in $Q_1(y_1) \cap Q_2(x_1)$ tending to infinity. In the latter case there is nothing more to say but in case $C^l(x_1)$ leaves $\overline{Q_1(y_1)}$, Theorem 4.5 and the fact that $\overline{Q_2(y_1)}$ is fixed point free leave only two alternatives. Either $C^l(x_1)$ tends to infinity in $Q_2(y_1) \cap Q_2(x_1)$ or $C^l(x_1)$ terminates on the vertical coordinate axis. Since $C^l(x_1)$ cannot intersect $C^l(y_1)$, the latter alternative can only occur if $C^l(y_1)$ terminates on the vertical coordinate axis as in Fig. 4.5(a). The three other curves are treated in similar fashion. The number of possibilities for each of these curves prevents a catalogue of all possibilities. In Fig. 4.6 below we combine Figs. 4.5 and 4.4 together with the observations above.

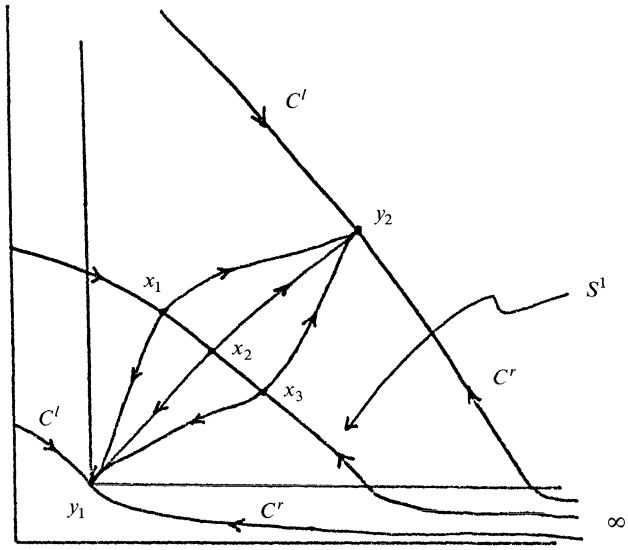


FIG. 4.6. *Figs.* 4.5(a) *and* 4.4 *are combined. The curves* $C^l(y_2)$ *and* $C^r(y_2)$ *are also featured. Note the curve* $S^1 = C^l(x_1) \cup C^r(x_1) \cup C^l(x_3) \cup C^r(x_3)$ *forms the boundary between the* $B(y_1)$ *and* $B(y_2)$.

Let S^1 denote the curve consisting of $C^l(x_1) \cup C^r(x_1) \cup C^l(x_3) \cup C^r(x_3) \cup \cdots \cup C^r(x_{2n+1})$ (see Fig. 4.6). S^1 is a C^1 monotone decreasing curve which disconnects R^2_+ into two components, a lower one, L^1, and an upper one, U^1. We will soon see that $B(y_1) = L^1$. First though we must address the second question raised above. Consider the lower left-hand component whose upper boundary is $C^l(y_2) \cup C^r(y_2)$. This region may be further partitioned into $4n+5$ separate regions formed by $C^l(y_1)$, $C^r(y_1)$, $C^l(y_2)$, $C^r(y_2)$ and the $C^+(x_i)$, $C^-(x_i)$ for odd i. Can there be a fixed point in any one of these open regions? We argue that there cannot be any fixed points in these regions. A formal proof would proceed case by case through each region, systematically using (E_+) and (E_-) of Theorem 2.5 and (E_l) and (E_r) of Theorem 4.5. We proceed less formally considering a sample case and leaving the remainder to the reader. Consider the open region, V, bounded by $C^l(x_1)$, a portion of the vertical coordinate axis, $C^l(y_2)$ and $C^+(x_1)$ (see Fig. 4.7 below). This region is mapped into itself by T. There can be no fixed points of T in $V \cap [x_1, y_2]$ by Theorem 2.5 (E_+). There can be no fixed points in $V \cap Q_2(x_1)$ by Theorem 4.5 (E_l). There can be no fixed points of T in $Q_2(y_2) \cap V$ by Theorem 4.5 (E_l). This exhausts the region. Indeed, one can show that the orbit of every point of V tends to y_2 by using the above arguments. The other regions can be treated in similar fashion; note that each is mapped into itself by T.

We show $L^1 = B(y_1)$ as follows. Each of the $(4n+5)$ regions are mapped into themselves by T. Corollary 4.4, the fact that each open region is fixed point free, and the fact that y_1 is the only stable fixed point in L^1 complete the argument.

We now state the main result of this section, completely describing the dynamics of an orientation preserving cooperative map in the case that A., B. and C. hold.

THEOREM 4.8. *Let* T *be an orientation preserving cooperative map satisfying* A., B. *and* C. *and assume* $B(x_m) \neq R^2_+$. *Then the set of fixed points is either infinite or odd in number and can be described as follows. The subset of stable fixed points is totally ordered* $0 < x_m \equiv y_1 < y_2 < \cdots < y_n \cdots$. *In each order interval* $[y_k, y_{k+1}]$, $k = 1, \cdots$, *there are an*

odd number of unrelated fixed points $x_1^k, x_2^k, \cdots, x_{2n_k+1}^k$ *lying on a* C^1 *monotone decreasing separatrix curve* S^k *and ordered from left to right on* S^k. *For each fixed point* x_i^k, $C^-(x_i^k)$ *and* $C^+(x_i^k)$ *tend to* y_k *and* y_{k+1} *respectively and are tangent at these points to the positive eigenvector of* $DT(y_k)$ *or* $DT(y_{k+1})$. *The odd indexed fixed points* x_{2i+1}^k *are saddle points and the even indexed fixed points are repellers.* $S^k = C^l(x_1) \cup C^r(x_1) \cup C^l(x_3) \cup C^r(x_3) \cup \cdots \cup C^l(x_{2n+1}) \cup C^r(x_{2n+1}) \cup \{x_2, x_4, \cdots, x_{2n}\}$ *and* S^k *separates* R_+^2 *into two components. If* L^k *denotes the lower left component of* R_+^2 *and* U^k *the upper right component of* R_+^2 *the boundary of which is* S^k *then* $B(y_1) = L^1$, $B(y_k) = L^k \cap U^{k-1}$, $k > 1$, *unless the sequence of stable fixed points terminates at* y_p. *In the latter case* $B(y_p) = U^{p-1}$.

See Fig. 4.8 below for a possible scenario.

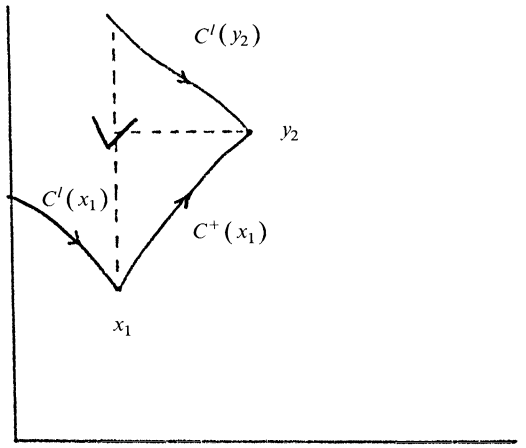


FIG. 4.7. *The region V in the case that* $C^l(x_1)$ *intersects the vertical coordinate axis.*

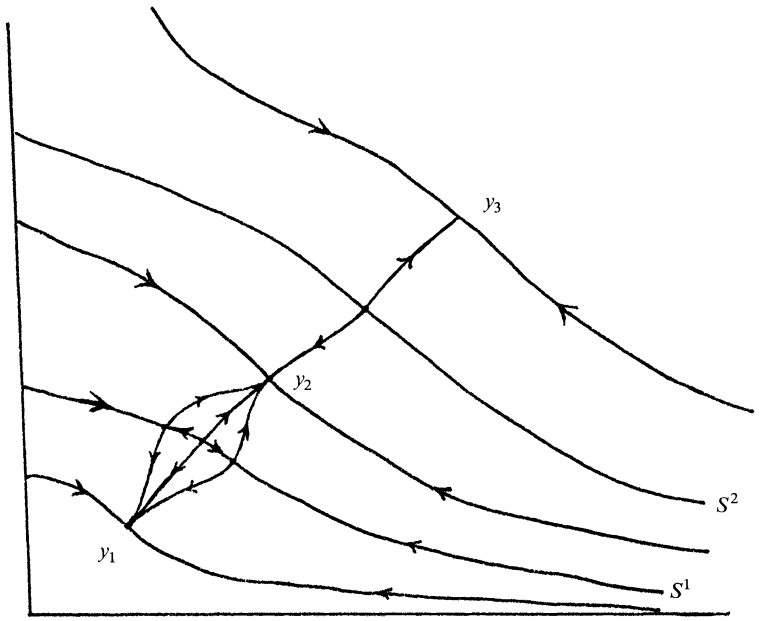


FIG. 4.8. *A phase portrait consistent with Theorem* 4.8. *There are three stable fixed points* y_1, y_2, y_3 *with separatrices* S^1 *and* S^2 *separating* $B(y_1)$, $B(y_2)$ *and* $B(y_3)$. *The interval* $[y_1, y_2]$ *contains three unstable fixed points on* S^1 *and* $[y_2, y_3]$ *contains one unstable fixed point on* S^2.

Theorem 4.8 is proved by successively employing Propositions 4.6 and 4.7 together with arguments using Theorem 2.5 (E_+), (E_-) and Theorem 4.5 (E_l) and (E_r) to insure regions bounded by the various curves are fixed point free. For example, the region lying above $C^l(y_2)$ in $Q_2(y_2)$ must be shown to be fixed point free as well as the region lying above $C^r(y_2)$ in $Q_4(y_2)$. Theorem 4.5 (E_l) and (E_r) respectively can be used to show that these regions are fixed point free. Now only $\overline{Q_1(y_2)}$ needs to be considered and Propositions 4.6 and 4.7 can be applied.

We end this section with the proofs of Proposition 4.6 and 4.7.

Proof of Proposition 4.6. Suppose $B^+(y_1) \neq \overline{Q_1(y_1)}$. Let $x \in S^+(y_1)$ and $\bar{x} = \lim_{n\to\infty} T^n x$ (the limit exists by C. and Corollary 4.4). Then $\bar{x} \in S^+(y_1)$, $T\bar{x} = \bar{x}$ and $\bar{x}$ is a saddle point. Observe that $\bar{x} \in Q_1(y_1)$ and $C^l(\bar{x}) \cap \overline{Q_1(y_1)}$ and $C^r(\bar{x}) \cap \overline{Q_1(y_1)}$ are contained in $S^+(y_1)$. If $C^l(\bar{x})$ has a limit point z in $\overline{Q_1(y_1)}$, then it is a repelling fixed point which lies in $S^+(y_1)$ and $C^l(\bar{x})$ is tangent at z to the eigenvector for $DT(z)$ corresponding to $\eta(z)$. A similar statement holds for $C^r(\bar{x})$. The first assertion of Proposition 4.6 is established by piecing together the curves $C^l(\bar{x})$ and $C^r(\bar{x})$ keeping in mind Proposition 2.3 (note that every ray $y_1 + th$, $h > 0$, intersects $S^+(y_1)$ in one point). $S^+(y_1)$ is C^1 because where a $C^r(\bar{x}_1)$ meets a $C^l(\bar{x}_2)$ at a repelling fixed point, the two curves are tangent to the same eigendirection by Theorem 4.5.

Now, either $S^+(y_1)$ intersects both boundary lines of $\overline{Q_1(y_1)}$, in which case $S^+(y_1)$ contains a finite number of fixed points by B. none of which can lie on the boundary lines, or $S^+(y_1)$ is unbounded. In either case the set of fixed points on $S^+(y_1)$ must lie discretely on $S^+(y_1)$, they cannot accumulate, and they must alternate between saddle and repeller. Now each such fixed point has $\rho > 1$ and so by Theorem 2.5 and C. each fixed point on $S^+(y_1)$ has a monotone invariant curve C^+ emanating from it and connecting the fixed point to a stable fixed point. In Proposition 4.7, we will show that all of the curves C^+ tend to the *same* fixed point. Since each of these curves C^+ is monotone, it would be impossible for the set of fixed points of T on $S^+(y_1)$ to be unbounded.

Proof of Proposition 4.7. In view of Theorem 2.5 and C., only the assertion that all the curves $C^+(x_i)$ are asymptotic to the same fixed point and the final assertion concerning $C^l(y_2)$ and $C^r(y_2)$ require proof. Consider a particular $x_i \in S^+(y_1)$ and the limit point z_i of $C^+(x_i)$ which exists by C. The fixed point z_i is stable and the curve $C^r(z_i)$ must, by Theorem 4.5, lie in $Q_4(z_i)$ and either (a) tends to infinity monotonely in $Q_1(y_1)$, (b) leaves $Q_1(y_1)$ be crossing the horizontal boundary or (c) has a limit point which is a fixed point and lies in $Q_1(y_1)$. We rule out the possibility (c) as follows. If $C^r(z_i)$ limits on a fixed point $u_i \in Q_1(y_1) \cap Q_4(z_i)$ then u_i is a repeller (Theorem 4.5) so consider $C^-(u_i)$. $C^-(u_i) \subset Q_3(u_i)$ cannot intersect $S^+(y_1)$ nor can it leave $Q_1(y_1)$. Moreover $C^-(u_i)$ can not have its limit fixed point on $S^+(y_1)$ because this limit point must be stable. It follows that the limit point of $C^-(u_i)$ lies outside $B^+(y_1) \cup S^+(y_1)$ but in $Q_1(y_1)$. Let this point be v_i, a stable fixed point, and consider $C^l(v_i)$. In Fig. 4.9 below we sketch the curves and fixed points discussed above. The curve $C^l(v_i)$ cannot cross any of the previously mentioned curves so it must have a limit point $w_i \in Q_2(v_i)$ below $C^+(x_i)$. Since w_i is a repeller and $C^+(w_i)$ cannot cross any of the curves previously mentioned, it must limit on a stable fixed point. We may clearly continue this reasoning producing an inward spiral $C^+, C^r, C^-, C^l, C^+, C^r, C^-, C^l, \cdots$ (see Fig. 4.9). The limit fixed points on these successive curves must accumulate. This violates B. Hence we see that possibility (c) cannot occur. Similar reasoning applies to $C^l(z_i)$: either it tends to infinity in $Q_1(y_1)$ or leaves $Q_1(y_1)$. See Fig. 4.10 below. But now, in view of Fig. 4.9, consider z_{i-1} and z_{i+1} the limit points of $C^+(x_{i-1})$ and $C^+(x_{i+1})$

respectively. These curves cannot cross $C^l(z_i)$ or $C^r(z_i)$. If $z_{i-1} \neq z_i$ then z_{i-1} must lie in the region bounded by $S^+(y_1)$, $C^+(x_i)$ and $C^l(z_i)$. Since $C^r(z_{i-1})$ cannot cross any of these curves, we have a contradiction to the fact that $C^r(z_{i-1})$ must be either unbounded or leave $Q_1(y_1)$. Thus we must have $z_{i-1} = z_i = z_{i+1}$.

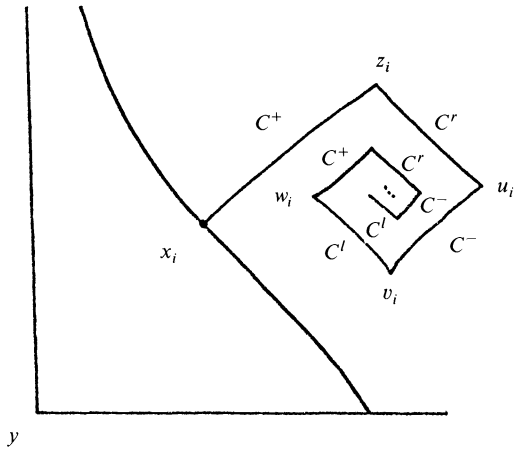


FIG. 4.9. *Inward spiral* C^+, C^r, C^-, C^l, $C^+, \cdots$.

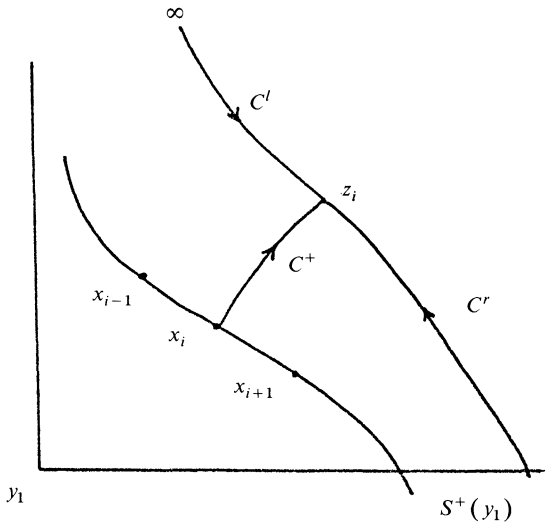


FIG. 4.10. *The sets* $C^l(z_i)$ *and* $C^r(z_i)$. *Where can* $C^l(z_{i+1})$ *and* $C^r(z_{i-1})$ *go?*

Acknowledgment. The author is grateful to the referee for pointing out a paper of Hirsch [16] in which Propositions K and L appear as well as other results pertinent to this study.

REFERENCES

[1] J. CARR, *Applications of Centre Manifold Theory*, Springer-Verlag, New York, 1981.

[2] W. COPPEL, *Stability and Asymptotic Behavior of Differential Equations*, D. C. Heath, Boston, 1965.

[3] J. K. HALE AND A. S. SOMOLINOS, *Competition for fluctuating nutrient*, J. Math. Biology, 18 (1983), pp. 255–280.

[4] P. HARTMAN, *Ordinary Differential Equations*, P. Hartman, Baltimore, 1973.

[5] M. W. HIRSCH, *Systems of differential equations which are competitive or cooperative*. I: *Limit sets*, this Journal, 13 (1982), pp. 167–179.

[6] _____, *Systems of differential equations which are competitive or cooperative*. II: *Convergence almost everywhere*, this Journal, 16 (1985), pp. 423–439.

[7] S. KARLIN AND H. M. TAYLOR, *A First Course in Stochastic Processes*, 2nd ed., Academic Press, New York, 1975.

[8] M. A. KRASNOSEL'SKII, *Translation Along Trajectories of Differential Equations*, A.M.S. Translation 19, Providence, RI, 1968.

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

[10] P. DE MOTTONI AND A. SCHIAFFINO, *Competition systems with periodic coefficients: A geometric approach*, J. Math. Biol., (1981), pp. 319–335.

[11] J. SELGRADE, *Asymptotic behavior of solutions to single loop positive feedback systems*, J. Differential Equations, 38 (1980), pp. 80–103.

[12] S. SMALE, *On the differential equations of species in competition*, J. Math. Biol., 3 (1976), pp. 5–7.

[13] H. L. SMITH, *Invariant curves for mappings*, this Journal, 17 (1986), pp. 1053–1067.

[14] R. S. VARGA, *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1962.

[15] W. J. LEONARD AND R. MAY, *Nonlinear aspects of competition between species*, SIAM J. Appl. Math., 29 (1975), pp. 243–275.

[16] M. W. HIRSCH, *Attractors for discrete time monotone dynamical systems in strongly ordered spaces*, Proc. Special Year in Geometry, Univ Maryland, College Park, 1983–1984.

SIAM J. MATH. ANAL.
Vol. 17, No. 6, November 1986

002

PERIODIC SOLUTIONS OF FORCED NONLINEAR SECOND ORDER EQUATIONS: SYMMETRY AND BIFURCATIONS*

M. FÜRKOTTER† AND H. M. RODRIGUES‡

Abstract. The authors are concerned with the equation $\ddot{u}+u=g(u,p)+\mu f(t)$, where p, μ are small parameters, f is an even, continuous π-periodic function, g is an odd smooth function of u, such that $g(u,p)=O(|pu|+|u^3|)$, as p and u go to zero. The main results are that, under certain conditions, the small 2π-periodic solutions maintain some symmetry properties of the forcing function $f(t)$, when $\mu\neq 0$. Some other interesting results describe the changes in the number of such solutions as p and μ vary in a small neighborhood of the origin. The authors use the approach of alternative problems.

Key words. periodic solutions, symmetry, bifurcation, nonlinear equations, small solutions

AMS(MOS) subject classifications. Primary 34A34, 34C15, 34C25

1. Introduction.

We are concerned with the equation

$$\ddot{u}+u=g(u,p)+\mu f(t) \tag{1.1}$$

where p, μ are small parameters, f is an even continuous periodic function and g is sufficiently smooth.

Our main results are that if g is odd in u and small near $(u,p)=(0,0)$, f is π-periodic and some conditions are satisfied, then the small 2π-periodic solutions of (1.1) maintain some symmetry properties of the forcing function $f(t)$, when $\mu\neq 0$. We also find the bifurcation curves and describe the changes of the number of such solutions as (p,μ) varies in a small neighborhood of the origin.

Hale–Rodrigues [3], [1] studying Duffing's equation, $\ddot{u}+u=pu-u^3+\mu\cos t$, showed that the only small 2π-periodic solutions are even functions of t, if $\mu\neq 0$. They also stated the same result for a general even forcing function with minimal periodic 2π under the condition $\int_0^{2\pi} f(s)\cos s\,ds\neq 0$.

Rodrigues–Vanderbauwhede [4] generalized this result for equations like (1.1), where f satisfies the former hypothesis and $g(u,p)=O(|pu|+u^2)$ as (u,p) goes to $(0,0)$. They also presented an abstract version for equations in Banach spaces. Vanderbauwhede [5], [6] also considers problems related to the above ones in an abstract form.

Many authors look for solutions of nonlinear equations in classes of functions defined by symmetry conditions. For instance, some of them consider equations similar to (1.1) with f 2π-periodic, and just prove that there are 2π-periodic solutions $u(t)$, which are even functions of t, or equivalently, $\dot{u}(0)=0$. But Hale, Vanderbauwhede and Rodrigues in the cited papers went further and proved that, under certain conditions, these are the only small 2π-periodic solutions.

Taking the same attitude of the last three authors, we study the small 2π-periodic solutions of (1.1), but assume that f has minimal period π. In this case the condition of Hale–Rodrigues, $\int_0^{2\pi} f(s)\cos s\,ds\neq 0$, no longer holds and in addition to the even solutions, solutions with other symmetries may arise. The main features of this paper

*Received by the editors August 21, 1984, and in revised form September 1, 1985. This research was supported in part by FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo, Brazil) under processo 80/0253-0.

†Instituto de Planejamento e Estudos Ambientais, UNESP, Presidente Prudente, Brazil.

‡Instituto de Ciências Matemáticas de São Carlos, USP, São Carlos, Brazil.

are to find a set of small 2π-periodic solutions of (1.1) and to prove that these are the only feasible solutions.

In §2, using the Lyapunov–Schmidt method, we show that symmetries in (1.1) imply symmetries in the solution of the auxiliary equation.

In §3, under the conditions

$$g(u,p)=pu+\alpha u^3+\cdots,\alpha\int_0^{2\pi}[(\mathscr{K}f)(t)]^2\cos 2t\,dt\neq 0,$$

we prove that the only small 2π-periodic solutions of (1.1) are such that either $u(t)$ or $u(t-\pi/2)$ is even in t, if $\mu\neq 0$, where $\mathscr{K}f$ indicates the π-periodic solutions of $\ddot{u}+u=f(t)$. As an example we analyse the equation $\ddot{u}+u=pu+u^3+\mu(1+\cos 2t)$.

However, our main results and the harder part of this work are in §4. If $f(t+\pi/2)=-f(t)$ the condition of §3 on $\mathscr{K}f$ is not satisfied. In this case $f(t-\pi/2)$ is even and $f(t\pm\pi/4)$ is odd in t. We show that, if a certain coefficient is not zero, the only small 2π-periodic solutions $u(t)$ of (1.1) are such that $u(t)$ or $u(t-\pi/2)$ is even or $u(t\pm\pi/4)$ is odd in t, if $\mu\neq 0$. This is stated in Theorem 4.2. We call special attention to Theorem 4.1, which was not easy to state and plays an important role in the proof of Theorems 4.2 and 4.3. The calculations indicate that the bifurcation equations are more degenerate when more symmetries are presented in (1.1). As an example we analyse the equation $\ddot{u}+u=pu+u^3+\mu\cos 2t$.

In §§3 and 4 we give, in Theorems 3.2 and 4.3 respectively, a complete description of the bifurcation curves and of the number of small 2π-periodic solutions, as (p,μ) varies in a small neighborhood of the origin.

It is not our aim to discuss all the possible cases, but at the end of §4 and below we give some indications of what can be expected when some of our assumptions are not satisfied, for example, when g is not odd and when f has other periods.

An application which can be reduced to (1.1) is the equation of a forced pendulum $\ddot{v}+(g/L)\sin v=\sigma f(wt)$ where σ is small, w is close to $w_0=\sqrt{g/L}$, f is π-periodic and we look for $2\pi/w$-periodic solutions. If we let $u(t)\overset{\text{def}}{=}v(t/w)$ and $w_0^2/w^2\overset{\text{def}}{=}1-p$ we get an equation like (1.1).

It is convenient to point out that the conditions to be verified in our main theorems usually do not involve hard calculations, because they can be computed as long as one knows the forcing function and a few terms of Taylor expansion of the nonlinearity.

Equation (2.5b) is important for the determination of the admissible phases. After this is done the problem is reduced to the analysis of the equation (2.5a) which will provide the "amplitude" r and the bifurcation diagram.

If $g(u,p)$ is odd in u and f is $2\pi/n$-periodic, even, and not odd harmonic, our conjecture is that, in general, there exists an integer $m=m(n)$, with $m(1)=0$, $m(n)>0$ if $n>1$, such that (2.5b) is given by

$$G(r,\phi,p,\mu)=\begin{cases} r^m\mu\sin n\phi(\rho_n+\cdots) & \text{if } n \text{ is odd,}\\ r^m\mu^2\sin n\phi(\rho_n+\cdots) & \text{if } n \text{ is even,}\end{cases}$$

where $\rho_n=\rho_n(f,g)$, ρ_n does not depend on ϕ and depends only on finitely many coefficients of the Taylor expansion of $g(u,0)$ around 0.

If $g(u,p)$ is odd in u, f is even and $f(t+\pi/n)=-f(t)$, our conjecture is that generically, there exists an integer $m=m(n)$, with $m(1)=0$, $m(n)>0$ if $n>1$, such that

equation (2.5b) is given by

$$G(r,\phi,p,\mu)=\begin{cases} r^m\mu\sin n\phi(\eta_n+\cdots) & \text{if } n \text{ is odd,}\\ r^m\mu^2\sin 2n\phi(\eta_n+\cdots) & \text{if } n \text{ is even,}\end{cases}$$

where $\eta_n=\eta_n(f,g)$, η_n does not depend on ϕ and depends only on finitely many coefficients of the Taylor expansion of $g(u,0)$ around 0.

The analysis of some examples indicates evidences that if g is odd in u and factor term is $\sin n\phi$ the the factor term in r is r^{n-1}, that is $m=n-1$.

If we do not assume that g is odd in u, a broader conjecture is that, generically, if f is even and $2\pi/n$-periodic then there exist integers m, p, q, which depend on n such that

$$G(r,\phi,p,\mu)=r^m\mu^p\sin q\phi(\xi_n+\cdots)$$

where ξ_n depends only on finitely many coefficients of the Taylor expansion of $g(u,0)$ around 0 and $m>1$ if $n>1$.

When ρ_n (η_n or ξ_n) does not vanish, we can discuss the existence of some special solutions.

In any of the above cases, for $n>1$, $r=0$ would give rise to the $2\pi/n$-periodic solution.

2. The auxiliary and the bifurcation equations. Consider the equation

$$\ddot{u}+u=g(u,p)+\mu f(t) \tag{1.1}$$

where (p,μ) varies in a small neighborhood of the origin, and the following hypotheses:

(A_1) f is a real π-periodic, even function, continuous on R.

(B_1) g is a C^∞ real function defined in a neighborhood of $(u,p)=(0,0)$, odd in u, and $g(u,p)=pu+\alpha u^3+O(|pu^3|+|u^5|)$ as (p,μ) goes to $(0,0)$. In fact it would be enough to assume that g is sufficiently smooth.

If $u(t)$ is a 2π-periodic solution of (1.1), then there are $r\in R$ and $\phi\in(-\pi/2,\pi/2]$, such that $u(t)=r\cos(t-\phi)+x(t)$, where

$$\int_0^{2\pi}x(t)\cos t\,dt=\int_0^{2\pi}x(t)\sin t\,dt=0.$$

If we let $u(t+\phi)\overset{\text{def}}{=}r\cos t+v(t)$, then $v(t)$ is a solution of

$$\begin{aligned}&\ddot{v}+v=g(r\cos t+v,\ p)+\mu f(t+\phi),\\ &\int_0^{2\pi}v(t)\cos t\,dt=\int_0^{2\pi}v(t)\sin t\,dt=0.\end{aligned} \tag{2.1}$$

Let $\mathscr{P}$ be the space of all 2π-periodic real functions, continuous on R, with the norm $\|w\|=\sup_{0\le t\le 2\pi}|w(t)|$, and let $\mathscr{P}^{(2)}$ be the space of all 2π-periodic real functions, with the second derivative continuous on R, with the norm $\|w\|=\sup\{|w^{(j)}(t)|,\ 0\le t\le 2\pi,\ j=0,1,2\}$.

On these spaces we consider the projection

$$(Pw)(t)\overset{\text{def}}{=}\frac{\cos t}{\pi}\int_0^{2\pi}w(s)\cos s\,ds+\frac{\sin t}{\pi}\int_0^{2\pi}w(s)\sin s\,ds. \tag{2.2}$$

The Fredholm alternative implies that the equation $\ddot{u}+u=h(t)$, with h in $\mathscr{P}$, has a solution in $\mathscr{P}^{(2)}$ if and only if $Ph=0$. Moreover, if $Ph=0$ then there exists a unique solution $u(t)$ in $\mathscr{P}^{(2)}$, of this equation, such that $Pu=0$. We indicate this solution by $\mathscr{K}h$. From the variation of constants formula, we obtain

$$\mathscr{K}h=(I-P)\left[-\cos(\cdot)\int_0^{(\cdot)}h(s)\sin s\,ds+\sin(\cdot)\int_0^{(\cdot)}h(s)\cos s\,ds\right]. \tag{2.3}$$

Following the usual procedure of the Lyapunov–Schmidt method (see [2]), the problem is reduced to that of finding v in $\mathscr{P}^{(2)}$ for the following system of equations,

$$\begin{aligned}&\text{(a) } v=\mathscr{K}(I-P)[g(r\cos(\cdot)+v,\ p)+\mu f(\cdot+\phi)],\\ &\text{(b) } P[g(r\cos(\cdot)+v,\ p)+\mu f(\cdot+\phi)]=0.\end{aligned} \tag{2.4}$$

The equations (a) and (b) are called the auxiliary and bifurcation equation, respectively. It follows from the implicit function theorem that (2.4a) has a unique small solution for (p,μ) in a small neighborhood of the origin. We denote this solution by $v^*(t)=v^*(r,\phi,p,\mu)(t)$. If we substitute in (2.4b), we obtain the following equations:

$$\begin{aligned}&\text{(a) } F(r,\phi,p,\mu)\overset{\text{def}}{=}\frac{1}{\pi}\int_0^{2\pi}g(r\cos s+v^*(r,\phi,p,\mu)(s),\ p)\cos s\,ds=0,\\ &\text{(b) } G(r,\phi,p,\mu)\overset{\text{def}}{=}\frac{1}{\pi}\int_0^{2\pi}g(r\cos s+v^*(r,\phi,p,\mu)(s),\ p)\sin s\,ds=0\end{aligned} \tag{2.5}$$

The following lemma gives information about some symmetries and estimates of v^*.

LEMMA 2.1. *If hypotheses* (A_1) *and* (B_1) *are satisfied, then the solution* v^* *of* (2.4a) *has the following properties*:

$$v^*(r,\phi,p,\mu)(t)=v^*(-r,\phi,p,\mu)(t-\pi), \tag{2.6}$$

$$v^*(r,\phi,p,\mu)(t)=-v^*(r,\phi,p,-\mu)(t-\pi), \tag{2.7}$$

$$v^*(0,\phi,p,\mu)(t)\ \textit{is } \pi\textit{-periodic in } t, \tag{2.8}$$

$$v^*(0,\phi,p,\mu)(t-\phi)\ \textit{is even in } t \textit{ and independent of } \phi, \tag{2.9}$$

$$v^*(r,0,p,\mu)(t),\ v^*(r,\pi/2,p,\mu)(t) \textit{ and } v^*(r,\phi,p,0)(t) \textit{ are even functions of } t, \tag{2.10}$$

$$v^*(0,\phi,p,\mu)=\mu\mathscr{K}f(\cdot+\phi)+O(|p\mu|+|\mu|^3) \textit{ as } (p,\mu) \textit{ goes to } (0,0), \tag{2.11}$$

$$v^*(r,\phi,p,\mu)=v^*(0,\phi,p,\mu)+rS(r,\phi,p,\mu) \textit{ where } S(r,\phi,p,\mu)=O(|r|^2+|\mu|) \textit{ as } (r,p,\mu) \textit{ goes to } (0,0,0). \tag{2.12}$$

Proof. Properties (2.6) to (2.10) follow essentially from the fact that the auxiliary equation is invariant under certain transformations.

To prove (2.11) it suffices to observe that

$$v^*(0,\phi,0,0)\equiv 0,\quad \frac{\partial v^*}{\partial\mu}(0,\phi,0,0)=\mathscr{K}f(\cdot+\phi),$$

$$\frac{\partial^2 v^*}{\partial\mu^2}(0,\phi,0,0)=0,\quad \frac{\partial v^*}{\partial p}(0,\phi,0,0)=0.$$

To get (2.12) we define

$$S(r,\phi,p,\mu) \overset{\text{def}}{=} r^{-1}\left[v^*(r,\phi,p,\mu)-v^*(0,\phi,p,\mu)\right] \quad \text{if } r\neq 0,$$

$$S(0,\phi,p,\mu) \overset{\text{def}}{=} \frac{\partial v^*}{\partial r}(0,\phi,p,\mu),$$

and observe that $S(r,\phi,p,0)=O(r^2)$. □

The next lemma is related to the bifurcation equation (2.5).

LEMMA 2.2. *Suppose hypotheses* (A_1), (B_1) *are satisfied. Then* F *and* G *given in* (2.5) *are odd in* r, *even in* μ *and* $G(r,\phi,p,0)\equiv 0$, *for* (r,p,μ) *in a small neighborhood of the origin.*

Proof. The first part follows from (2.6) and the second part is a consequence of (2.7) and (2.10). □

LEMMA 2.3. *Suppose hypotheses* (A_1), (B_1) *are satisfied. Then, for* (r,p,μ) *in a small neighborhood of the origin,* $G(r,\phi,p,\mu)=r\mu^2\sin 2\phi(\rho+\cdots)$, *where*

$$\rho=-3\alpha(2\pi)^{-1}\int_0^{2\pi}(\mathscr{K}f(\tau))^2\cos 2\tau\,d\tau$$

and $\cdots$ *indicates terms of order* $O(|p|+|\mu|+|r|)$, *uniformly with respect to* ϕ, *as* (r,p,μ) *goes to* $(0,0,0)$.

Proof. From (2.10) it follows that

$$G(r,0,p,\mu)\equiv 0 \quad \text{and} \quad G\left(r,\frac{\pi}{2},p,\mu\right)\equiv 0. \tag{2.13}$$

Applying (2.13) and Lemma 2.2 we obtain $G(r,\phi,p,\mu)=r\mu^2\sin 2\phi H(r,\phi,p,\mu)$, where H is a smooth function.

From (2.12) and (B_1) it follows that

$$G(r,\phi,p,\mu)=\pi^{-1}\left[3\alpha r^2\int_0^{2\pi}\cos^2 s\,v^*(s)\sin s\,ds+3\alpha r\int_0^{2\pi}\cos s\left(v^*(s)\right)^2\sin s\,ds\right.$$
$$\left.+\alpha\int_0^{2\pi}\left(v^*(s)\right)^3\sin s\,ds+o\left(r^2+\mu^2\right)\right].$$

By (2.11) and (2.12),

$$\frac{\partial^3 G}{\partial\mu^2\partial r}(0,\phi,0,0)=-\frac{3\alpha}{\pi}\sin 2\phi\int_0^{2\pi}(\mathscr{K}f(\tau))^2\cos 2\tau\,d\tau.$$

From this expression we obtain the value of ρ and this completes the proof of our lemma. □

3. The non odd-harmonic case. In this section we give conditions on the nonlinear term and on the forcing function in such a way that the only small 2π-periodic solutions $u(t)$ of (1.1) are such that, either $u(t)$ or $u(t-\pi/2)$ is an even function of t. We also descibe the bifurcation diagram. As an example we consider the equation $\ddot{u}+u=pu+u^3+\mu(1+\cos 2t)$.

THEOREM 3.1. *Suppose hypotheses* (A_1), (B_1) *are satisfied and that* $\alpha\int_0^{2\pi}(\mathscr{K}f(\tau))^2\cos 2\tau\,d\tau\neq 0$, *where* $\mathscr{K}f$ *is the* π*-periodic solution of* $\ddot{u}+u=f(t)$. *Then the only small* 2π*-periodic solutions* $u(t)$ *of* (1.1) *are such that either* $u(t)$ *or* $u(t-\pi/2)$ *is an even function of* t, *for* (p,μ) *in a small neighborhood of the origin and* $\mu\neq 0$.

Proof. By Lemma 2.3, $G(r,\phi,p,\mu)=r\mu^2\sin 2\phi(\rho+\cdots)$. We see that $r=0$ solves $G(r,\phi,p,\mu)=0$. Since in this case, $u(t)=v^*(0,\phi,p,\mu)(t-\phi)$, it follows from (2.9) that $u(t)$ is even in t. For $\mu\neq 0$, since $\rho\neq 0$, the only left possibility is that $\sin 2\phi=0$. This implies $\phi=0,\ \pi/2$, because we have assumed that ϕ is in $(-\pi/2,\pi/2]$. Recalling that $u(t)=r\cos(t-\phi)+v^*(r,\phi,p,\mu)(t-\phi)$, it follows from (2.10) that for $\phi=0$, $u(t)$ is even in t. For $\phi=\pi/2$, we have, from (2.7) that $v^*(r,\pi/2,p,\mu)(t-\pi)=v^*(-r,\pi/2,p,\mu)(t)$. Then $u(t-\pi/2)=r\cos(t-\pi)+v^*(-r,\pi/2,p,\mu)(t)$ is an even function of t, as a consequence of (2.10). □

We should point out that the above theorem still does not prove the existence of 2π-periodic solutions, because the first bifurcation equation was not solved yet. It only states that if some solution exists it must have the mentioned properties.

Now we turn to (2.5a). By using hypothesis (B_1), we let $g(u,p)\overset{\text{def}}{=}pu+\alpha u^3+pu^3h_1(u,p)+u^5h_2(u)$, where h_1, h_2 are smooth functions. If we write

$$h_1(r\cos t+v^*(r,\phi,p,\mu)(t),\ p)=h_1(v^*(0,\phi,p,\mu)(t),\ p)+O(r),$$

$$h_2(r\cos t+v^*(r,\phi,p,\mu)(t))=h_2(v^*(0,\phi,p,\mu)(t))+O(r)$$

and use (2.12), (2.8), (2.11), we can show that, $F(r,\phi,p,\mu)=r[p+\frac{3}{4}\alpha r^2+3\alpha\lambda\mu^2+o(|p|+r^2+\mu^2)]$, where

$$\lambda=\lambda(\phi)=\frac{1}{2\pi}\left[\int_0^{2\pi}(\mathscr{K}f(\tau))^2\,d\tau+\cos 2\phi\int_0^{2\pi}((\mathscr{K}f)(\tau))^2\cos 2\tau\,d\tau\right]. \tag{3.1}$$

It is clear that $r=0$ solves $F(r,\phi,p,\mu)=0$ and we already know that this will give rise to a π-periodic solution of (1.1).

Let $J(r,\phi,p,\mu)\overset{\text{def}}{=}r^{-1}F(r,\phi,p,\mu)$ if $r\neq 0$ and $J(0,\phi,p,\mu)\overset{\text{def}}{=}F_r(0,\phi,p,\mu)$. In order to find the multiple roots of $J=0$, we must solve the system

$$\begin{aligned} J(r,\phi,p,\mu)&=p+\frac{3}{4}\alpha r^2+3\alpha\lambda\mu^2+\cdots=0,\\ J_r(r,\phi,p,\mu)&=\frac{3}{2}\alpha r+\cdots=0, \end{aligned} \tag{3.2}$$

where $\cdots$ indicates higher order terms.

Since $\det(\partial J,J_r)/\partial(p,r))=\frac{3}{2}\alpha$ for $r=p=\mu=0$ from the implicit function theorem, it follows from (3.2) that for $\alpha\neq 0$, p and r can be solved for as functions of μ in a neighborhood of the origin, for $\phi=0$ and $\phi=\pi/2$. But Lemma 2.2 implies that $F(r,\phi,p,\mu)$ is an odd function of r. This shows that $J_r(0,\phi,p,\mu)\equiv 0$. Thus $r\equiv 0$ is the function of μ given above. The functions $p=p(\mu)$ can be found by solving $J(0,\phi,p,\mu)=p+3\alpha\lambda\mu^2+o(|p|+|\mu|^2)=0$, and we obtain $p=-3\alpha\lambda\mu^2+O(\mu^4)$.

If we assume that $\alpha\int_0^{2\pi}\cos 2\tau[(\mathscr{K}f)(\tau)]^2\,d\tau\neq 0$, an analysis of (3.1) shows that $\lambda_1\overset{\text{def}}{=}\lambda(0)$ and $\lambda_2\overset{\text{def}}{=}\lambda(\pi/2)$ are nonzero and distinct. Then we have two bifurcation curves, Γ_1 and Γ_2, for $\phi=0$ and $\phi=\pi/2$, respectively, given by

$$\begin{aligned} \Gamma_1:&\quad p=-3\alpha\lambda_1\mu^2+O(\mu^4),\\ \Gamma_2:&\quad p=-3\alpha\lambda_2\mu^2+O(\mu^4). \end{aligned}$$

Using the fact that J is quadratic in r, we can prove that when we cross each bifurcation curve we gain or lose two solutions. In Fig. 3.1 we show the bifurcation diagram with the number of 2π-periodic solutions of (1.1), in a neighborhood of the origin, when λ_1, λ_2 and α are positive. Thus, we have proved the following.

THEOREM 3.2. *Suppose* (A_1), (B_1) *are satisfied and that* $\alpha\int_0^{2\pi}\cos 2\tau[(\mathscr{K}f)(\tau)]^2 d\tau \neq 0$, *where* $\mathscr{K}f$ *is the* π*-periodic solution of* $\ddot{u}+u=f(t)$. *Then there exist two bifurcation curves* Γ_1 *and* Γ_2, *in the plane* (p,μ), *given by* $p=-3\alpha\lambda_1\mu^2+O(\mu^4)$ *and* $p=-3\alpha\lambda_2\mu^2+O(\mu^4)$, *respectively, where* λ_1 *and* λ_2 *are obtained from* (3.1), *for* $\phi=0$, *and* $\phi=\pi/2$, *respectively. These curves divide a neighborhood of the origin into three regions, as shown in Fig.* 3.1. *In* S_1 *there is only one* 2π*-periodic solution which is in fact* π*-periodic. In* S_2 *there are two solutions with minimum period* 2π *and one with period* π. *In* S_3 *there are four solutions with minimum period* 2π *and one with period* π. *In* Γ_1 *we have only one* 2π*-periodic solution, which is in fact* π*-periodic; and in* Γ_2 *we have three* 2π*-periodic solutions, but one of them is* π*-periodic. All of the above information is concerned with small solutions, for* $\mu\neq 0$.

Example 3.1. Let us consider (1.1) with $f(t)=1+\cos 2t$ and $g(u,p)=pu+u^3$. In this case we have Duffing's equation. The calculation shows that $\rho=1$, $\lambda_1=\lambda(0)=\frac{13}{18}$, $\lambda_2=\lambda(\pi/2)=\frac{25}{18}$ and the bifurcation curves are given by $\Gamma_1: p=-\frac{13}{6}\mu^2+O(\mu^4)$, $\Gamma_2: p=-\frac{25}{6}\mu^2+O(\mu^4)$ and the picture is shown in Fig. 3.1.

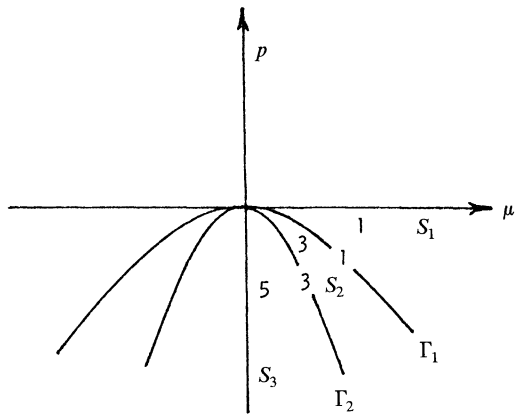


FIG. 3.1

4. The odd-harmonic case. This section contains our main results and is the most difficult part of this paper. The reason for the difficulty is that more symmetries are present in (1.1) and as a consequence, the bifurcation equation is more degenerate.

In this part we will assume more restrictive conditions on the forcing function and on the nonlinearity.

(A_2) $f: R\to R$ is an even continuous function and $f(t+\pi/2)=-f(t)$, for all t in R.

(B_2) g is a real C^∞ function, defined in a neighborhood of the origin, $g(u,p)=pu+\alpha u^3+\nu u^5+O(|pu^3|+|u^7|)$ and g is odd in u.

Besides the information given by Lemma 2.1, we have in this case the following properties of the solution of the auxiliary equation (2.4a).

LEMMA 4.1. *Suppose* (A_2), (B_2) *are satisfied. Then, for all* t *in* R,

$$v^*(0,\phi,p,\mu)(t+\pi/2)=-v^*(0,\phi,p,\mu)(t). \tag{4.1}$$

Furthermore, $v^*(r,\pi/4,p,\mu)(t-\pi/2)$ *is odd in* t.

The proof follows the same ideas as for Lemma 2.1.

Remark. Besides the above information, we need in this section the following properties which follow easily from Lemma 2.1 under the assumptions (A_1), (B_1).

$$\frac{\partial v^*}{\partial r}(0,\phi,p,\mu)(t+\pi) = -\frac{\partial v^*}{\partial r}(0,\phi,p,\mu)(t), \tag{4.2}$$

$$\frac{\partial^2 v^*}{\partial r^2}(0,\phi,p,\mu)(t) \text{ is } \pi\text{-periodic in } t. \tag{4.3}$$

It was not easy to discover the representation stated in the next theorem. It plays an important role in the proof of our main results.

THEOREM 4.1. *Suppose* (B_1) *and* (B_2) *are satisfied. Then* $(\partial v^*/\partial r)(0,\phi,p,\mu)(t-\phi) = a(t)\cos(t-\phi)+b(t)\sin(t-\phi)$, *where* $a=a(p,\mu)$, $b=b(p,\mu)$ *are* $\pi/2$*-periodic functions, independent of* ϕ *and with mean value zero. Furthermore* $a=a(t)$ *and* $b=b(t)$ *are, respectively, even and odd in* t.

Proof. Let $M_\phi \overset{\text{def}}{=} \{ y(\cdot,\phi)\in\mathscr{P}^{(2)} : y(t,\phi)=a(t)\cos(t-\phi)+b(t)\sin(t-\phi)$, where a and b are $C^{(2)}$, $\pi/2$-periodic functions, with mean value zero, a is even and b is odd$\}$.

Next, we prove that M_ϕ is a closed subspace of $\mathscr{P}^{(2)}$. Let $y_n(t,\phi)=a_n(t)\cos(t-\phi)+b_n(t)\sin(t-\phi)$ be such that $y_n(\cdot,\phi)\to y(\cdot,\phi)$ in $\mathscr{P}^{(2)}$. An easy calculation shows that

$$\begin{aligned} a_n(t) &= y_n(t,\phi)\cos(t-\phi) - y_n(t+\pi/2,\phi)\sin(t-\phi), \\ b_n(t) &= y_n(t,\phi)\sin(t-\phi) + y_n(t+\pi/2,\phi)\cos(t-\phi). \end{aligned} \tag{4.4}$$

Then (a_n) and (b_n) are convergent in $\mathscr{P}^{(2)}$. Let a and b be their limits.

Relations (4.4) imply that $y(t,\phi)=a(t)\cos(t-\phi)+b(t)\sin(t-\phi)$. The other properties which define M_ϕ are easily verified.

Now we observe that $(\partial v^*/\partial r)(0,\phi,p,\mu)(t-\phi)$ is a solution of the equation

$$\ddot{x}+x = (I-P)\left[\frac{\partial g}{\partial u}\big(v^*(0,\phi,p,\mu)(\cdot-\phi),\ p\big)(\cos(\cdot-\phi)+x)\right].$$

We define $H:\mathscr{P}^{(2)}\to\mathscr{P}$ by

$$Hy = \mathscr{K}(I-P)\left[\frac{\partial g}{\partial u}\big(v^*(0,\phi,p,\mu)(\cdot-\phi),\ p\big)(\cos(\cdot-\phi)+y)\right].$$

Our next purpose is to prove that H leaves M_ϕ invariant. As a first step we claim that if $\alpha(t)$, $\beta(t)$ are $C^{(2)}$, $\pi/2$-periodic functions, $\alpha(t)$ is even, $\beta(t)$ is odd and if we assume that β has mean value zero, then

$$z(\cdot,\phi) \overset{\text{def}}{=} (I-P)(\alpha(\cdot)\cos(\cdot-\phi)+\beta(\cdot)\sin(\cdot-\phi))$$

is an element of M_ϕ. This follows from

$$z(t,\phi) = \left[\alpha(t) - (2\pi)^{-1}\int_0^{2\pi}\alpha(s)\,ds\right]\cos(t-\phi)+\beta(t)\sin(t-\phi).$$

Using the above claim, (B_1), Lemma 2.1 and Lemma 4.1, we can prove that

$$(I-P)\left[\frac{\partial g}{\partial u}\big(v^*(0,\phi,p,\mu)(\cdot-\phi),\ p\big)(\cos(\cdot-\phi)+y(\cdot-\phi))\right]$$

belongs to M_ϕ if $y(\cdot-\phi)$ is in M_ϕ.

Now, we prove that if $y(\cdot,\phi)$ is in M_ϕ, the $\mathscr{K}y(\cdot,\phi)$ is in M_ϕ. Since $\mathscr{K}$ is linear we will prove it for $y(t,\phi)=a(t)\cos(t-\phi)$ and the other part follows in a similar way.

By computing $\mathscr{K}y(\cdot,\phi)$, we obtain $\mathscr{K}y(\cdot,\phi)(t)=A(t)\cos(t-\phi)+B(t)\sin(t-\phi)$, where

$$A(t)=\frac{1}{2}\Bigg\{\Bigg[-\cos 2t\int_0^t a(s)\sin 2s\,ds-(2\pi)^{-1}\int_0^{2\pi}\cos 2s\,ds\int_0^s a(u)\sin 2u\,du$$

$$-\frac{1}{2\pi}\cos 2t\int_0^{2\pi}ds\int_0^s a(u)\sin 2u\,du\Bigg]$$

$$+\Bigg[\sin 2t\int_0^t a(s)\cos 2s\,ds-\frac{1}{2\pi}\int_0^{2\pi}\sin 2s\,ds\int_0^s a(u)\cos 2u\,du\Bigg]\Bigg\}.$$

$$B(t)=\frac{1}{2}\Bigg\{\int_0^t a(s)\,ds-\Bigg[\sin 2t\int_0^t a(s)\sin 2s\,ds-\frac{1}{2\pi}\sin 2t\int_0^{2\pi}ds\int_0^s a(u)\sin 2u\,du\Bigg]$$

$$-\cos 2t\int_0^t a(s)\cos 2s\,ds\Bigg\}.$$

With the identity

$$\frac{1}{2\pi}\int_0^{\pi}ds\int_0^s a(u)\sin 2u\,du=\int_0^{\pi/2}a(s)\sin 2s\,ds,$$

we prove that $A(t)$ and $B(t)$ are $\pi/2$-periodic. The remaining properties, which define M_ϕ, follow in a natural way. This concludes the proof that $\mathscr{K}y(\cdot,\phi)$ belongs to M_ϕ and so H leaves M_ϕ invariant.

Moreover, H is a uniform contraction with respect to (p,μ) in a neighborhood of the origin and ϕ in $(-\pi/2\pi/2]$.

From the contraction fixed point principle it follows that H has a unique fixed point in a small neighborhood of the origin in $\mathscr{P}^{(2)}$. Since M_ϕ is itself a Banach space and it is invariant under H, it follows that the fixed point must be in M_ϕ and this completes the proof that $(\partial v^*/\partial r)(0,\phi,p,\mu)(t-\phi)$ has the desired form. □

LEMMA 4.2. *Suppose hypotheses* (A_2), (B_2) *are satisfied. Then for* (r,p,μ) *in a small neighborhood of the origin*

$$G(r,\phi,p,\mu)=r^3\mu^2\sin 4\phi(\eta+\cdots), \tag{4.5}$$

where

$$\eta=\frac{1}{12\pi}\Bigg\{-27\alpha^2\int_0^{2\pi}\Big[\mathscr{K}(I-P)\Big((\mathscr{K}f(\cdot))^2\cos(\cdot)\Big)\Big](s)\cos 3s\,ds \tag{4.6}$$

$$-27\alpha^2\int_0^{2\pi}(\mathscr{K}f)(\tau)\mathscr{K}[(\mathscr{K}f)(\cdot)\cos 2(\cdot)](s)\cos 2s\,ds$$

$$+27\alpha^2\int_0^{2\pi}(\mathscr{K}f)(\tau)\mathscr{K}[(\mathscr{K}f)(\cdot)\sin 2(\cdot)](s)\sin 2s\,ds$$

$$+\Big[\frac{9}{16}\alpha^2-15\nu\Big]\int_0^{2\pi}[\mathscr{K}f(\tau)]^2\cos 4\tau\,d\tau\Bigg\},$$

$\mathscr{K}$ *is given by* (2.3) *and* $\cdots$ *indicates terms of order* $O(|p|+|\mu|+|r|)$, *uniformly on* ϕ, *as* (r,p,μ) *goes to* $(0,0,0)$.

Proof. From the last statement of Lemma 4.1, and (2.10), it follows that

$$G(r,\phi,p,\mu)\equiv 0 \quad \text{for } \phi=0,\ \pi/2,\ \pi/4. \tag{4.7}$$

Now we claim that

$$\frac{\partial G}{\partial r}(0,\phi,p,\mu)\equiv 0. \tag{4.8}$$

In fact, from Theorem 4.1,

$$\begin{aligned}\frac{\partial G}{\partial r}(0,\phi,p,\mu)&=\frac{1}{\pi}\int_0^{2\pi} g_u(v^*(0,\phi,p,\mu)(s-\phi),\ p)\frac{\partial v^*}{\partial r}(0,\phi,p,\mu)(s-\phi)\sin(s-\phi)\,ds\\&=\frac{1}{\pi}\int_0^{2\pi} g_u(v^*(0,\phi,p,\mu)(s-\phi),\ p)[a(s)\cos(s-\phi)+b(s)\sin(s-\phi)]\sin(s-\phi)\,ds\\&=\frac{1}{2\pi}\int_0^{2\pi} g_u(v^*(0,\phi,p,\mu)(s-\phi),\ p)(a(s)\sin 2(s-\phi)-b(s)\cos 2(s-\phi))\,ds\\&\quad+\frac{1}{2\pi}\int_0^{2\pi} g_u(v^*(0,\phi,p,\mu)(s-\phi),\ p)b(s)\,ds.\end{aligned}$$

Since $g_u(u,p)$ is even in u, from (4.1) it follows that $g_u(v^*(0,\phi,p,\mu)(s-\phi),\ p)$ is $\pi/2$-periodic in s. If we add to this the fact that a and b are $\pi/2$-periodic and that $\sin 2(s-\phi)$ and $\cos 2(s-\phi)$ change sign, when we change s by $s+\pi/2$, we conclude that the last but one integral vanishes. To prove that the last integral vanishes we use (2.9) and the fact that b is odd. This completes the proof of (4.8).

From Lemma 2.2, $G(r,\phi,p,\mu)$ is odd in r, even in μ and $G(r,\phi,p,0)\equiv 0$. If we use this information with (4.7) and (4.8), we can prove that there is a smooth function $S(r,\phi,p,\mu)$ such that $G(r,\phi,p,\mu)=r^3\mu^2\sin 4\phi S(r,\phi,p,\mu)$.

Since $g(u,0)=\alpha u^3+\nu u^5+O(u^7)$, we obtain

$$\begin{aligned}\frac{\partial^5}{\partial\mu^2\partial r^3}G(0,\phi,0,0)=\pi^{-1}\Big\{&18\alpha\int_0^{2\pi}\frac{\partial 3v^*}{\partial\mu^2\partial r}(0,\phi,0,0)(s)\cos^2 s\sin s\,ds\\&+36\alpha\int_0^{2\pi}\frac{\partial v^*}{\partial\mu}(0,\phi,0,0)(s)\frac{\partial^3 v^*}{\partial\mu\partial r^2}(0,\phi,0,0)(s)\sin s\cos s\,ds\\&+6\alpha\int_0^{2\pi}\left[\frac{\partial v^*}{\partial\mu}(0,\phi,0,0)(s)\right]^2\frac{\partial^3 v^*}{\partial r^3}(0,\phi,0,0)(s)\sin s\,ds\\&+120\nu\int_0^{2\pi}\left[\frac{\partial v^*}{\partial\mu}(0,\phi,0,0)(s)\right]^2\cos^3 s\sin s\,ds\Big\}.\end{aligned} \tag{4.9}$$

But

$$\frac{\partial^3 v^*}{\partial\mu^2\partial r}(0,\phi,0,0)=6\alpha\mathscr{K}(I-P)\big[(\mathscr{K}f(\cdot+\phi))^2\cos(\cdot)\big], \tag{4.10}$$

$$\frac{\partial v^*}{\partial\mu}(0,\phi,0,0)=\mathscr{K}f(\cdot+\phi), \tag{4.11}$$

$$\frac{\partial^3 v^*}{\partial\mu\partial r^2}(0,\phi,0,0)=6\alpha\mathscr{K}(I-P)\big[\mathscr{K}f(\cdot+\phi)\cos^2(\cdot)\big], \tag{4.12}$$

$$\frac{\partial^3 v^*}{\partial r^3}(0,\phi,0,0)=-\frac{3}{16}\alpha\cos 3(\cdot). \tag{4.13}$$

If we substitute (4.10)–(4.13) in (4.9), use the change $s=\tau-\phi$ in the integrals and take into account that $(\mathscr{K}(I-P)f)(t-\phi)=(\mathscr{K}(I-P)f(\cdot-\phi))(t)$ and $(\mathscr{K}f)(t-\phi)=(\mathscr{K}f(\cdot-\phi))(t)$, we obtain

$$\frac{\partial^5}{\partial\mu^2\partial r^3}G(0,\phi,0,0)=\frac{1}{\pi}\sin 4\phi\Big\{-27\alpha^2\int_0^{2\pi}\mathscr{K}(I-P)\big[(\mathscr{K}f(\cdot))^2\cos(\cdot)\big](\tau)\cos 3\tau\,d\tau$$
$$-27\alpha^2\int_0^{2\pi}(\mathscr{K}f)(\tau)\mathscr{K}\big[(\mathscr{K}f)(\cdot)\cos 2(\cdot)\big](\tau)\cos 2\tau\,d\tau$$
$$+27\alpha^2\int_0^{2\pi}(\mathscr{K}f)(\tau)\mathscr{K}\big[(\mathscr{K}f)(\cdot)\sin 2(\cdot)\big](\tau)\sin 2\tau\,d\tau$$
$$+\Big(\frac{9}{16}\alpha^2-15\nu\Big)\int_0^{2\pi}((\mathscr{K}f)(\tau))^2\cos 4\tau\,d\tau\Big\}.$$

This completes the proof of the lemma and gives the expression for η. □

Remark 4.1. It should be pointed out that the expression (4.6), which defines η, can be computed as long as we know the forcing function and the coefficients α and ν respectively of third and fifth order, of the Taylor expansion of $g(u,0)$ in u, around 0.

The next theorem is our main result. It claims that, under certain conditions, the small 2π-periodic solutions of (1.1) maintain some symmetry properties of the forcing function.

THEOREM 4.2. *Suppose hypotheses* (A_2), (B_2) *are satisfied and* $\eta\neq 0$, *where* η *is given by* (4.6). *Then the only small* 2π*-periodic solutions of* (1.1) *are such that* $u(t)$ *or* $u(t-\pi/2)$ *is even or* $u(t\pm\pi/4)$ *is odd in* t *for* (ρ,μ) *small and* $\mu\neq 0$.

Proof. From Lemma 4.2, $G(r,\phi,p,\mu)=r^3\mu^2\sin 4\phi(\eta+\cdots)$. We see that $r=0$ solves $G(r,\phi,p,\mu)=0$ and we already know that it will give rise to a π-periodic solution $u(t)$ of (1.1). Since $\eta\neq 0$, for $\mu\neq 0$ the only possibility left is that $\sin 4\phi=0$. This implies $\phi=0$, $\pi/2$, $\pm\pi/4$, because we have assumed that ϕ is in $(-\pi/2,\pi/2]$.

As in §3, Theorem 3.1, we can show that for $\phi=0$, the solution $u(t)$ is even in t and for $\phi=\pi/2$, $u(t-\pi/2)$ is even in t.

For $\phi=\pi/4$ since $u(t)=r\cos(t-\pi/4)+v^*(r,\pi/4,p,\mu)(t-\pi/4)$, from the last statement of Lemma 4.1 it follows that $u(t-\pi/4)$ is odd in t. The case $\phi=-\pi/4$ is similar. This completes the proof of our theorem. □

Now we turn to the analysis of the first bifurcation equation.

If we proceed as in the proof of Theorem 3.2, we see that bifurcation curves will be obtained by solving $J(0,\phi,p,\mu)=0$. The next lemma helps us to understand that equation.

LEMMA 4.3. *Suppose* (A_2), (B_2) *are satisfied. Then* $J(0,\phi,p,\mu)$ *is independent of* ϕ.

Proof. Using the definition of J,

$$J(0,\phi,p,\mu)=\frac{1}{\pi}\int_0^{2\pi}g_u\big(v^*(0,\phi,p,\mu)(s),\ p\big)\Big(\cos s+\frac{\partial v^*}{\partial r}(0,\phi,p,\mu)(s)\Big)\cos s\,ds.$$

If we let $s=\tau-\phi$ and proceed as in the proof of (4.8), by using (4.1) and Theorem 4.1 we obtain

$$J(0,\phi,p,\mu)=\frac{1}{2\pi}\Big(\int_0^{2\pi}g_u\big(v^*(0,\phi,p,\mu)(\tau-\phi),\ p\big)\,d\tau\Big)$$
$$+\int_0^{2\pi}g_u\big(v^*(0,\phi,p,\mu)(\tau-\phi)a(\tau)\,d\tau\big),$$

where $a(\tau)$ is given as in Theorem 4.1. Using (2.9) and Theorem 4.1, we see that $J(0,\phi,p,\mu)$ is independent of ϕ. □

The next theorem is very interesting and it describes the changes of the number of the small 2π-periodic solutions of (1.1) as (p,μ) crosses the bifurcation curve.

THEOREM 4.3. *Suppose* (A_2), (B_2) *are satisfied and* η, *given by* (4.6), *is nonzero. Then there exists a unique bifurcation curve* Γ, *given by* $p=-3\alpha\lambda\mu^2+O(\mu^4)$, *where* $\lambda=(2\pi)^{-1}\int_0^{2\pi}(\mathscr{K}f(s))^2\,ds$ *and* $\mathscr{K}f$ *is the* π*-periodic solution of* $\ddot{u}+u=f(t)$. *The curve* Γ *divides a neighborhood of the origin, in the plane* (p,μ) *into two regions* S_1, S_2, *as shown in Fig.* 4.1 *for* $\alpha,\lambda>0$. *For* (p,μ) *in* S_1 *there is a unique* 2π*-periodic solution of* (1.1), *which is in fact* π*-periodic. In* S_2 *there are nine* 2π*-periodic solutions; only one of them is* π*-periodic. In* Γ *there is a unique* 2π*-periodic solution of* (1.1), *which is in fact* π*-periodic.*

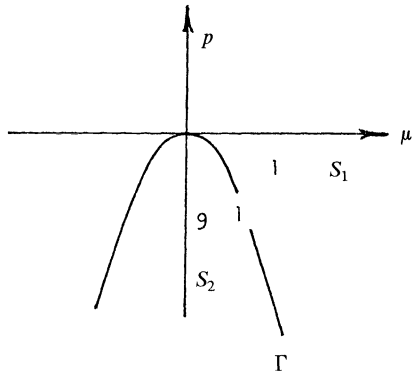


FIG. 4.1

All of the above information is concerned with small solutions of (1.1), *for* $\mu\neq 0$.

Proof. We proceed as in the proof of Theorem 3.2, by solving the equation

$$J(0,\phi,p,\mu)=p+3\alpha\lambda\mu^2+\cdots=0$$

where λ is given as in (3.1).

From (A_2) it follows that $\mathscr{K}f(\tau+\pi/2)=-\mathscr{K}f(\tau)$, and then

$$\lambda=\lambda(\phi)=\frac{1}{2\pi}\left[\int_0^{2\pi}(\mathscr{K}f(\tau))^2\,d\tau+\cos 2\phi\int_0^{2\pi}(\mathscr{K}f(\tau))^2\cos 2\tau\,d\tau\right]$$

$$=\frac{1}{2\pi}\int_0^{2\pi}(\mathscr{K}f(\tau))^2\,d\tau,$$

which shows that λ is independent of ϕ. Since, from Lemma 4.3, $J(0,\phi,p,\mu)$ is independent of ϕ, we conclude by implicit function theorem, that the bifurcation curve is the same for $\phi=0$, $\pi/2$, $\pm\pi/4$ and it is given by $p=-3\alpha\lambda\mu^2+O(\mu^4)$.

Since $J(r,\phi,p,\mu)$ is quadratic in r we see that for each phase, $\phi=0$, $\pi/2$, $\pm\pi/4$, when we cross Γ, we gain or lose two solutions, $u(t)=r\cos(t-\phi)+v^*(r,\phi,p,\mu)(t-\phi)$, of (1.1) of minimum period 2π, besides the π-periodic one. This completes the proof of our theorem. □

The bifurcation is not so degenerate as it appears. In fact the above proof shows that, besides the π-periodic solution, for each fixed phase $\phi=0$, $\pi/2$, $\pm\pi/4$, we have a quadratic bifurcation in r.

Example 4.1. Consider the equation $\ddot{u}+u=pu+u^3+\mu\cos 2t$. In this case $\eta=-\frac{11}{8}$, $\lambda=\frac{1}{18}$, all the assumptions of Theorem 4.2 and Theorem 4.3 are satisfied and Γ is given by $p=-\frac{1}{6}\mu^2+O(\mu^4)$.

Example 4.2. Consider the equation of a forced pendulum $\ddot{v}+(g/L)\sin v=\sigma\cos 2w\tau$, where w varies near $w_0=\sqrt{g/L}$ and σ is small. With the changes $u(t)\overset{\text{def}}{=}v(t/w)$, $w_0^2/w^2\overset{\text{def}}{=}1-p$ and $\mu\overset{\text{def}}{=}\sigma/w^2$ the above equation is reduced to $\ddot{u}+u=g(u,p)+\mu\cos 2t$, where, $g(u,p)=pu+(u^3/3!)-(u^5/5!)+O(|pu^3|+|u^7|)$. In this case $\eta=-\frac{1}{32}$, $\lambda=\frac{1}{18}$, all the assumptions of Theorem 4.2 and Theorem 4.3 are satisfied and Γ is given by $p=-\frac{1}{36}\mu^2+O(\mu^4)$.

5. Final remarks. As we said in the Introduction, we do not intend to exhaust all the possible cases, but we give below some indications about what can be expected when some of our assumptions are not satisfied. Some preliminary calculations show that the ideas of our work can be used to solve other cases not included here.

1. Suppose that g is not odd in u, but $g(u,p)=pu+\alpha u^2+O(|pu^2|+|u^3|)$. Some calculations indicate that the bifurcation equations are given by

$$F(r,\phi,p,\mu)=r\left[p+\tfrac{5}{6}\alpha^2r^2+\alpha\lambda\mu+\cdots\right]=0,$$
$$G(r,\phi,p,\mu)=r\mu\sin 2\phi(\xi+\cdots)=0,$$

where

$$\lambda=\int_0^{2\pi}\mathscr{K}f(s)\,ds+\cos 2\phi\int_0^{2\pi}\mathscr{K}f(s)\cos 2s\,ds,$$
$$\xi=-\frac{\alpha}{\pi}\int_0^{2\pi}\mathscr{K}f(s)\cos 2s\,ds.$$

Thus, if $\xi\neq 0$ the only small 2π-periodic solutions $u(t)$ of (1.1) must satisfy either $u(t)$ is even in t or $u(t-\pi/2)$ is even in t, for $\mu\neq 0$.

The analysis of the first bifurcation equation furnishes the bifurcation curves.

2. Our approach should help to explain the case when the forcing function is even and $(2\pi/n)$-periodic, where n is a positive integer. For instance, some calculations for the example $\ddot{u}+u=pu+u^3+\mu\cos 3t$, show that the bifurcation equations are given by

$$F(r,\phi,p,\mu)=r\left[p+\tfrac{3}{4}r^2+\tfrac{3}{128}\mu^2+\cdots\right]=0,$$
$$G(r,\phi,p,\mu)=r^2\mu\sin 3\phi\left(\tfrac{3}{32}+\cdots\right)=0$$

and the only small 2π-periodic solutions $u(t)$ must satisfy either $u(t)$ is even or $u(t\pm\pi/3)$ is even in t.

Acknowledgment. We are grateful to a referee for valuable comments, in particular for a remark about the conjecture presented in the Introduction.

REFERENCES

[1] S. N. CHOW AND J. K. HALE, *Methods of BifurcationTheory*, Springer-Verlag, London–Heidelberg–Berlin–Tokyo, 1982.

[2] J. K. HALE, *Ordinary Differential Equations*, Krieger, New York, 1980.

[3] J. K. HALE AND H. M. RODRIGUES, *Bifurcation in the Duffing equation with independent parameters*. II, Proc. Royal Soc. Edinburgh A, 79 (1977), pp. 317–326.

[4] H. M. RODRIGUES AND A. VANDERBAUWHEDE, *Symmetric perturbations of nonlinear equations: symmetry of small solutions*, Nonlinear Anal. TMA, 2 (1978), pp. 27–46.

[5] A. VANDERBAUWHEDE, *Local Bifurcation and Symmetry*, Instituut voor Theoretische Mechanica, Rijksuniversiteit Gent, Belgium, 1980.

[6] _____, *Local Bifurcation and Symmetry*, Pitman Advanced Publishing Program, Boston-London-Milbourne, 1982.

SIAM J. MATH. ANAL.
Vol. 17, No. 6, November 1986

MULTIPLICITY OF SOLUTIONS OF NONLINEAR BOUNDARY VALUE PROBLEMS*

D. C. HART[†§], A. C. LAZER[‡] AND P. J. MCKENNA[†]

Abstract. Sharp results for the number of solutions of a one-dimensional nonlinear Neumann boundary value problem are given, in terms of the range of its linearization, and the projection of the source term onto the principle eigenfunction.

Key words. nonlinear Neumann problems, jumping nonlinearities, multiple solutions, bifurcations, scaling, perturbation, phase plane

AMS(MOS) subject classifications. Primary 34B15; secondary 35J65

1. There have been many investigations in recent years of nonlinear boundary value problems for equations of the form

$$\Delta u + g(u) = h(x) \tag{1}$$

under assumptions on the behavior of g at infinity:

$$\lim_{x \to -\infty} \frac{g(x)}{x} = a \quad \text{and} \quad \lim_{x \to +\infty} \frac{g(x)}{x} = b, \tag{2}$$

called asymptotically sublinear, linear, or superlinear if a and b are zero, finite, or infinite. Let $\lambda_1 < \lambda_2 \leqq \lambda_3 \cdots$ denote the eigenvalues of the linear problem

$$\Delta v + \lambda v = 0 \tag{3}$$

with, for example, Dirichlet boundary conditions. One may then investigate the effect of relationships between a, b and the $\{\lambda_i\}$. We suppose $a < b$ (so that g is what Fučik [8] terms a "jumping nonlinearity"); the cases $a < b$ and $b < a$ are equivalent under change of variables. If $\lambda < a \leqq g' \leqq b < \lambda_{n+1}$ for some n, then there is a unique solution of the Dirichlet problem (or the Neumann problem) for (1), for any smooth g and h; see, for example [5, Chap. 3].

Amann and Hess [1] and Dancer [6] have shown that if we decompose h as $s\phi_1 + h_1$, where ϕ_1 is the positive normalized eigenfunction for λ_1 and h_1 is orthogonal to ϕ_1 (in L^2), then there exists $S_0 = S_0(h_1)$ such that no solution exists for $s < S_0$, but (at least) two solutions exist for any $s > S_0$, if $a < \lambda_1 < b$. If also $b < \lambda_2$ and $g \in C^2$, with $g'' > 0$, then Berger and Podolak [4] have shown that there are exactly two solutions for $s > S_0$ (and exactly one at $s = S_0$); this extends to an earlier result of Ambrosetti and Prodi [2].

* Received by the editors July 17, 1984, and in revised form March 29, 1985.
† Department of Mathematics, University of Florida, Gainesville, Florida 32611.
‡ Department of Mathematics and Computer Sciences, University of Miami, Coral Gables, Florida 33124.
§ Present address, Department of Mathematics, University of Cincinnati, Cincinnati, Ohio 45267.

In [12], Lazer and McKenna conjecture that whenever $a<\lambda_1$, $\lambda_n<b<\lambda_{n+1}$ and h is as above, then for sufficiently large s there are at least $2n$ distinct solutions. Some partial results are given in [7], [10], [14], [17]. They have proved this for the one-dimensional case in [13]. The present authors have extended that result in [9], to show that if $a\in(\lambda_k,\lambda_{k+1})$ and $b\in(\lambda_n,\lambda_{n+1})$, for some integers $n>k\geqq 1$ (and a technical "nonresonance" condition on a,b is satisfied), then the number of solutions of

$$u''+g(u)=s\sin t+h_1(t), \tag{4a}$$

$$u(0)=u(\pi)=0, \tag{4b}$$

for large positive s, plus the number of solutions for large negative s, is at least $2(n-k+1)$.

In this note we discuss the corresponding one-dimensional Neumann problem,

$$u''+g(u)=s+h(t), \tag{5a}$$

$$u'(0)=0=u'(\pi). \tag{5b}$$

The eigenvalues of the corresponding linear problem (3) are then just the squares of integers. We verify the conjecture, and show that the estimate for the number of solutions is sharp in the asymptotically linear or sublinear cases, for any smooth h.

We assume without further mention a nondegeneracy condition, in the statement of our theorems: that g' does not vanish on an interval. If this condition were violated, the theorems would remain true if the phrase "any s" were replaced throughout by "any s such that $g(u)-s$ does not vanish on an interval." Our main result is the following.

THEOREM 1. *Suppose* $g\in C^1(R)$, *that* $\lim_{u\to-\infty}g'(u)=a$, $\lim_{u\to+\infty}g'(u)=b$, *where* $b\in((n-1)^2,n^2)$ *for some integer* $n\geqq 1$. *If* $a<0$, *then for any* $h\in C^1([0,\pi])$, *there exists* $S_0>0$ *such that for any* $s>S_0$, (5) *has exactly* $2n$ *solutions. If* $a\in((k-1)^2,k^2)$, *for some integer* k, $0<k<n$, *and* $a^{-1/2}+b^{-1/2}$ *is not twice the reciprocal of an integer, then for any* $h\in C^1([0,\pi])$, *there exists* $S_+>0$ *and* $S_-<0$ *such that the number of solutions of* (5) *for any* $s<S_-$ *plus the number of solutions of* (5) *for any* $s>S^+$ *is exactly* $2(n-k+1)$.

Note that we include the equilibrium solutions in our count. Our method is to use scaling arguments to reduce the asymptotically linear case to a perturbation of the autonomous piecewise linear case, which we analyze by phase plane methods.

For the superlinear case, we have the following.

THEOREM 2. *Suppose* $g\in C^1(R)$, *that* $\lim_{u\to-\infty}g'(u)=a$, *for some* $a\in[-\infty,\infty)$, *and that* $\lim_{u\to+\infty}g'(u)=+\infty$. *Then for any positive integer* n *there exists* S_n *such that for* $s>S_n$, *there exists* $\eta>0$ *such that if* $h\in C^0([0,\pi])$ *and* $\|h\|_{C^0}<\eta$, *then* (5) *has at least* $2n$ *solutions. Further, there exist* $S_-<0$ *such that* $s<S_-$ *implies, for small enough* h, *at least* $2k$ *solutions when* $a\in((k-1)^2,k^2)$.

If we do not assume that $\lim_{u\to+\infty}g'(u)$ exists, but only retain the idea that $g'(u)$ increases from negative to positive as u increases, we still have the following.

THEOREM 3. *Let* $g\in C^1(R)$, *and suppose there exist numbers* $R_a<R_b$ *and* $a<0<b$ *such that* $u<R_a$ *implies that* $g'(u)<a$, *and* $u>R_b$ *implies that* $g'(u)>b$. *Assume that* $b>(n-1)^2$ *for some integer* $n\geqq 1$. *Then there exist* $S>0$ *such that for any* $s>S$, *there exists* $\eta>0$ *such that* $\|h\|_{C^0}<\eta$ *implies that* (5) *has at least* $2n$ *solutions.*

Finally, we note that these results would not be altered by the introduction of a "sufficiently small" dissipative term, $\varepsilon u'$, in (5). The plan of the papers is as follows: in §2, we establish the lower bounds on the number of solutions as given in Theorems 1, 2,

and 3. In §3, we establish the upper bound on the number of solutions as given in Theorem 1 for piecewise linear autonomous equations. In §4, we show that we can reduce the general situations of Theorem 1 to perturbations of the problem treated in §3.

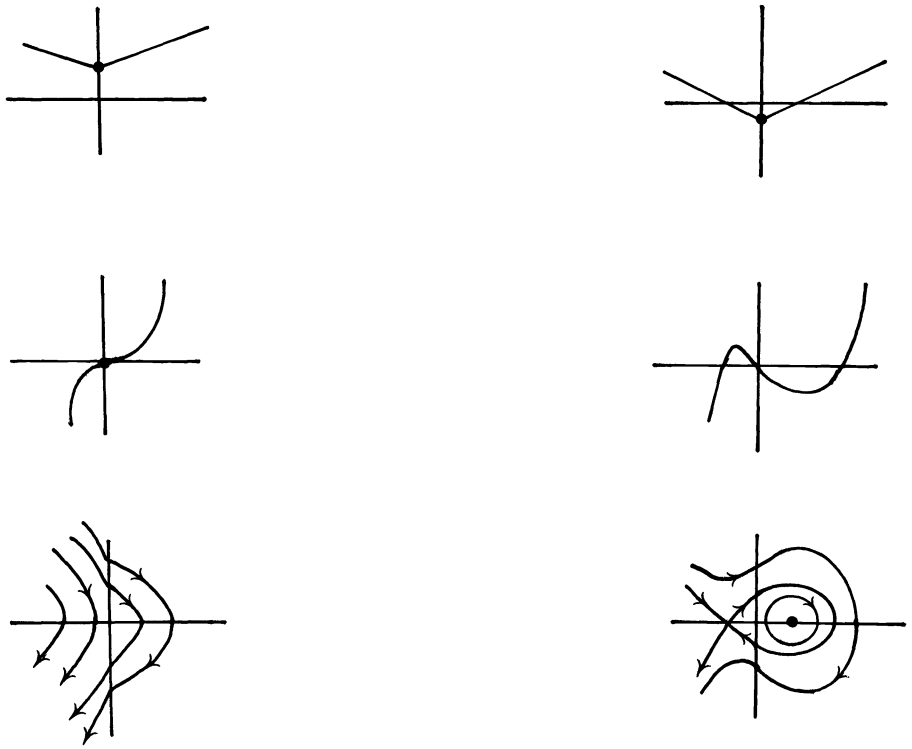
2. Solutions to the autonomous equation (h = constant)

$$u''+g(u)=s, \tag{6a}$$

$$u'(0)=0=u'(\pi), \tag{6b}$$

lie in the level curves of the conserved "energy" $E=\frac{1}{2}(u')^2+V(u)$, where the "potential" $V(u)=\int^u [g(t)-s]\,dt$ (see for example [3] or [11]). Solutions of the Neumann problem (6) are simply segments of 2π-periodic trajectories. Any periodic solution lies in a "well" of the potential, that is, a neighborhood of a local minimum of V. These minima, in general, are just the solutions of $g(u)=s$ such that $g'(u)>0$ (see Fig. 1b). At such an equilibrium point C (a "center"), the period tends to $2\pi(g'(c))^{-1/2}$, the period of the linearization of the vector field

$$u'=v, \qquad v'=s-g(u) \tag{7}$$



(a) $s<0$. (b) $s>0$.

FIG. 1

at the constant solution $u=C$. Note that if g' is monotone there can be only one such well, so just one center C. Further, by choice of s we may drive C toward positive or negative infinity (for a,b not zero), and so control $g'(C)$. Note also that the nondegeneracy hypothesis, that g' does not vanish on an interval, implies that $g(u)-s$ does not vanish on an interval but at discrete points; this is the only use we make of this hypothesis.

Suppose that the hypotheses of Theorem 3 are satisfied; for some numbers $R_a<R_b$, $g'(u)<a<0$ if $u<R_a$, and $g'(u)>b>0$ if $u>R_b$, where $b>(n-1)^2$ for some positive integer n. Then there must be at exactly two equilibria for s large, since $g(u)\to\infty$ as $u\to\pm\infty$; let C be the larger, and S the smaller. Then as the energy E increases from $V(C)$ to $V(S)$, the periods go from $2\pi(g'(C))^{-1/2}$ to infinity; hence if s is such that $g'(C)>(n-1)^2$, there exists a $(2\pi/k)$-periodic trajectory $(u_k(t),v_k(t))$ of (7) for each integer k between 1 and $n-1$. Hence the Neumann problem (6) has at least $2n$ solutions (including the constant solutions $u\equiv C$ and $u\equiv S$), obtained by requiring that $v_k(0)=0$. Let $\phi(t,u_0,v_0)=(u(t,u_0,v_0),\ v(t,u_0,v_0))$ be the time-t map (the flow) for the solution of (7) with initial conditions $u(0)=u_0$, $v(0)=v_0$. Then for some neighborhood of $u_k(0)$, where $u_k(t)$ is a solution of the Neumann problem (6), $\phi(\pi)$ carries the u-axis across itself, with a zero at $u_k(\pi)$. If h is C^0-small, then the time-π map for (5) must also carry the u-axis across itself, and thus (5) also has a solution (near the solution $u_k(t)$ of (6)). The persistence of the constant solutions follows from the inverse function theorem, in the standard way [5]. This proves Theorem 3, and Theorem 2 for $a<0$.

We now treat the case $a>0$ of Theorem 2. If $\lim_{u\to-\infty}g'(u)=a\in((k-1)^2,k^2)$ for an integer $k\geqq 1$, and $\lim_{u\to+\infty}g'(u)=+\infty$, then $g(u)=s$ has a unique solution $C=C(s)$ for large positive s, and also for large negative s; all solutions are then periodic. Choose S_n such that $g'(C)>n^2$ for $s>S_n$; then the periods of (6) approach some value less than $2\pi/n$, as the solutions approach C. Since n may be arbitrary, we need only show that the greatest period is bounded below. Let $x=u-C$ and $f(x)=g(x+C)-s$, so that (6a) becomes $x''+f(x)=0$, where $f(0)=0$. Introduce polar coordinates by $x=r\cos\theta$, $y=x'/\omega=r\sin\theta$, where we treat ω as a free parameter. Then

$$\theta'=(xy'-yx')/r^2=-\omega\left[1-\left(1-\frac{g(x)}{\omega^2x}\right)\left(\frac{x}{r}\right)^2\right]. \tag{8}$$

Let the right side of (8) define $-F(\omega,r,\theta)$. Take $\omega=\sqrt{a}$, and consider a segment of a trajectory with $\theta(0)=5\pi/4$, $\theta(T)=3\pi/4$, so x/r is bounded below, and take R so large that $|1-f(x)/ax|<1$, for $x\leqq x(0)=-R/\sqrt{2}$. Then $|\theta'|\leqq\sqrt{a}(1+|1-f(x)/ax|\cos^2\theta)\leqq 2\sqrt{a}$ and the time T elapsed on this trajectory segment exceeds $\pi/4\sqrt{a}$. The range of periods for (6a) contains $(2\pi/n,\pi/4\sqrt{a})$, so as n increases, the number of solutions of (6) increases without bound.

We now take $S_-<0$ so that for $s<S_-$, $|1-f(x)/ax|<\varepsilon$, for all $x\leqq 0$. We want to show that $k-1$ 2π-periodic solutions exist when $(k-1)^2<\lim_{u\to-\infty}g'(u)<k^2$. The periods of trajectories near C approach $2\pi(g'(C))^{-1/2}$, so exceed $2\pi/k$ for ε sufficiently small. On the other hand, we claim that as the energy goes to infinity, the periods become less than $\pi/(k-1)$ since (i) the time-elapsed for $x<0$ is less than $\pi/(k-1)$, and (ii) the time elapsed for $x>0$ is going to zero. The first assertion may be seen from (8), with $\omega=\sqrt{a}$ again; we have $|\theta'|=\sqrt{a}(1-\varepsilon\cos^2\theta)$, and if ε is so small that $k-1<\sqrt{a}(1-\varepsilon)$, then $T<\pi/(k-1)$.

We now show (ii), that the time T^+ for $x>0$ may be made arbitrarily small. Note that $T^+=2\int_0^{\pi/2}d\theta/F(\omega,r,\theta)$, where F was defined in equation (8). Thus if g is increased (keeping $g(0)=0$) then F is also increased, so that T^+ is decreased. Now

define x_j by requiring $g'(x)>(j+1)^2$ for all $x\geqq x_j$, and define $\theta_j(r)=\cos^{-1}(x_j/r)$ for $r\geqq x_j$. If g is monotone, then $x_{k-1}<0<x_k$ (since $g'(0)\in((k-1)^2,k^2)$; let j^* denote the least $j\geqq -1$ such that $x_{j+1}\geqq 0$). Notice that $\theta_{j+1}<\theta_j$ and $\theta_j\to 0$ as $j\to\infty$, for fixed r, and that $\theta_j(r)\to\pi/2$ as $r\to\infty$, for any j. Given $\varepsilon>0$, choose $n>j_*$ such that $\pi/n<\varepsilon/2$; then choose R so large that for any $r>R$, $\pi/2-\theta_n<\varepsilon/4$. Since $|f'(0)-a|<\varepsilon$, the time for θ to go from $\pi/2$ to θ_n is less than $\varepsilon/4a$, on a trajectory with $r(0)>R$; the time from θ_n to 0 is less than $\varepsilon/4$; and thus $T^+<\varepsilon$.

Finally, a C^0-small nonautonomous perturbation h is treated as in the case $a<0$ above. This completes the proof of Theorem 2.

3. Consider now the piecewise linear function $g(u)=au$ for $u<0$, $g(u)=bu$ for $u>0$; we write $g(u)=bu^+-au^-$. Then (6) becomes

$$u''+bu^+-au^-=s, \tag{9a}$$

$$u'(0)=0=u'(\pi). \tag{9b}$$

We consider two cases, (i) $a<0$ and $(n-1)^2<b<n^2$, and (ii) $(k-1)^2<a<k^2$, $(n-1)^2<b<n^2$.

In the former ($a<0$) case, the graphs of $g(u)-s$, $V_s(u)$, and the phase portrait are as in Fig. 1. There are no periodic solutions if $s<0$. If $s>0$, the period of the strictly positive orbits is $2\pi/\sqrt{b}$; we show below that as the energy increases from zero, the period tends continuously and monotonically to infinity. Hence there are $n-1$ nonstationary 2π-periodic solutions if $n-1<\sqrt{b}<n$, and the Neumann problem (9) has exactly $2n$ solutions.

In the case $\sqrt{a}\in(k-1,k)$ and $\sqrt{b}\in(n-1,n)$ for some integrals $n>k\geqq 1$, the situation is as depicted in Fig. 2. All solutions are periodic and consist of solutions of the two linear problems, pieced together along $x=0$. If $s=0$, all have period $\pi(a^{-1/2}+b^{-1/2})$. There are positive solutions for $s>0$, which have period $2\pi b^{-1/2}$, and negative solutions for $s<0$ with period $2\pi a^{-1/2}$. Using (8), one finds the period of a solution of both signs passing through $(u,u')=(0,I)$, $I>0$, to be $2((\pi-\theta_a)a^{-1/2}+\theta_b b^{-1/2})$, where $\tan\theta_a=-Ia^{1/2}/s$, and $\tan\theta_b=-Ib^{1/2}/s$ (θ_a and θ_b are in $[0,\pi]$). The period is then easily seen to be continuous and strictly monotone (increasing with I for $s>0$, decreasing for $s<0$), with range $(2\pi b^{-1/2},\pi(a^{-1/2}+b^{-1/2}))$ for $s>0$ and $(\pi(a^{-1/2}+b^{-1/2}),2\pi a^{-1/2})$ for $s<0$. Since we have assumed that $a^{-1/2}+b^{-1/2}$ is not twice the reciprocal of an integer, the number of solutions of (9) for s positive, plus the number for s negative, is therefore exactly $2(n-k+1)$, as claimed.

4. We treat the asymptotically linear problem (5) as a perturbation of the autonomous piecewise-linear problem (9), via rescaling. Let $u=|s|v$, so that (9a) becomes

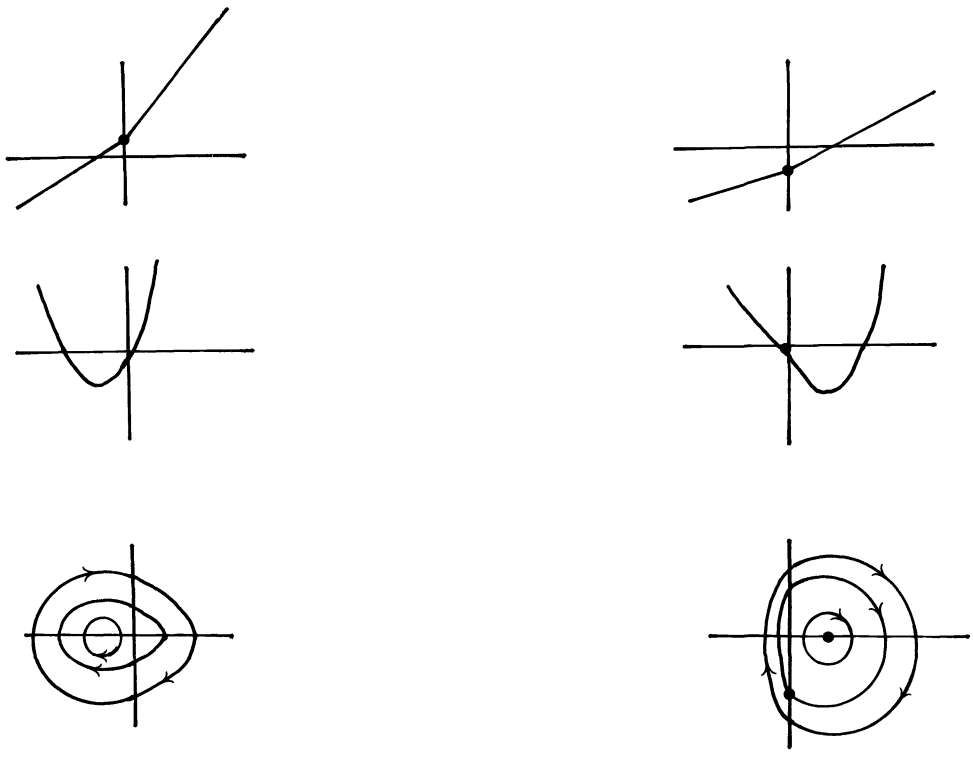
$$v''+bv^+-av^-=\pm 1 \tag{10}$$

and (5a) becomes

$$v''+bv^+-av^-\big(bv^+-av^--g(|s|v)/|s|\big)=\pm 1+h(x)/|s|. \tag{11}$$

We want to show that (11) differs from (10) by a C^1-small term for $|s|$ large, since then their solutions are C^1-close. If $\theta_E(t)$ is the angular coordinate on the solution of (9) with energy E and with $\theta(0)=\pi$, then near solutions of the Neumann problem, $\theta_E(\pi)$ carries the negative x-axis to a smooth curve which is differentiably transverse to the positive x-axis. Hence any map C^1-close to θ_E does the same. We therefore see that for

any given a, b and s, any C^1 perturbation of (9) has the same number of solutions. (We remark that the permissible size of such a perturbation decreases as a or b approach an eigenvalue n^2, for fixed s, or as b is increased.)



(a) $s<0$. (b) $s>0$.

FIG. 2. $(k-1)^2<a<k^2$, $(n-1)^2<b<n^2$.

If v is a solution of (10), the trajectory (v,v') in phase space is bounded away from the center $(C/|s|,0)$ if s is bounded above, and its size is bounded above if s is bounded below. We require $|s|$ to be so large that $\|h\|_{C^1}<\eta/2|s|$, and likewise $\|bv^+-av^--g(|s|v)/|s|\|_{C^1}<\eta/2|s|$, for any v in an annulus $r_1<\|(v,v')-(C,0)\|<r_2$. We will treat $s>0$; the other case is similar. For $v>0$, we then have $bv^+-av^--g(sv)/s=v(b-g(sv)/sv)$. Since v is bounded (for $s>S_0$, say, where S_0 is such that $g'(C)$ exceeds n^2 by $\frac{1}{2}(b-n^2)$ for all $s>S_0$), and $g(u)/u\to b$ as $u\to\infty$, it follows that $\|v(b-qg(sv)/sv)\|_{C^1}\to 0$ as $s\to\infty$. Likewise for $v<0$, $v(a-g(sv)/sv)\to 0$. This gives C^0-smallness (for large positive s); C^1-smallness follows, as $v>0$ implies $(bv^+-av^--g(sv)/s)'=b-g'(sv)\to 0$ as $s\to\infty$, similarly for $v<0$. Thus (5) has exactly as many solutions as (9), and the proof of Theorem 1 is completed.

The term $h(x)$ is of course no problem above, under the assumption that $\lim_{x\to+\infty}g'(x)$ exists and is finite. Under the assumptions of Theorem 2, however, the $|s|$ required above increases as η decreases, as b increases; so if $g'(x)$ is unbounded, then for fixed $\|h\|_{C^1}$, we may be forced to choose s large to arrange that $\|h\|_{C^1}/|s|<\eta$, but then find $g'(C)$ (the "b" above) so large that η must be still smaller.

REFERENCES

[1] H. AMANN AND P. HESS, *A multiplicity result for a class of elliptic boundary value problems*, Proc. Roy. Soc. Edinburgh, 84A (1979), pp. 145–151.

[2] A. AMBROSETTI AND G. PRODI, *On the inversion of some differentiable mappings with singularities between Banach spaces*, Annali Mat. Pura Appl., Ser IV, 3 (1972), pp. 231–247.

[3] V. I. ARNOLD, *Ordinary Differential Equations*, MIT Press, Cambridge, MA, 1973.

[4] M. S. BERGER AND E. PODOLAK, *On the solutions of a nonlinear Dirichlet Problem*, Indiana Univ. Math. J., 24 (1975), pp. 837–846.

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

[6] E. N. DANCER, *On the ranges of certain weakly nonlinear elliptic partial differential equations*, J. Math. Pures Appl., 57 (1978), pp. 351–366.

[7] _____, *Degenerate critical points, homotopy indices, and Morse inequalities*, preprint.

[8] S. FUČIK, *Boundary value problems with jumping nonlinearities*, Casopis Pest. Mat., 101 (1976), pp. 69–87.

[9] D. C. HART, A. C. LAZER AND P. J. MCKENNA, *Multiplicity of solutions of nonlinear boundary value problems with jumping nonlinearities*, J. Differential Equations, 59 (1985), pp. 266–281.

[10] H. HOFER, *Variational and topological methods in partiallly ordered Banach spaces*, Math. Ann., 261 (1982), pp. 493–514.

[11] D. W. JORDAN AND P. SMITH, *Nonlinear Ordinary Differential Equations*, Oxford Univ. Press, Oxford, 1977.

[12] A. C. LAZER AND P. J. MCKENNA, *On the number of solutions of a nonlinear Dirichlet problem*, J. Math. Anal. Appl., 84 (1981), pp. 282–294.

[13] _____, *On a conjecture related to the number of solutions of a nonlinear Dirichlet problem*, Proc. Roy. Soc. Edinburgh, to appear.

[14] _____, *Multiplicity results for a class of semilinear elliptic and parabolic boundary value problems*, J. Math. Anal. Appl., to appear.

[15] _____, *Multiple solutions of boundary value problems with nonlinearities crossing higher eigenvalues*, to appear.

[16] W. S. LOUD, *Periodic solutions of* $x'' + cx' + g(x) = \varepsilon f(t)$, Memoirs Amer. Math. Soc., 31, 1959.

[17] S. SOLIMINI, *Existence of a third solution for a class of B.V.P. with jumping nonlinearities*, Nonlinear Anal., 7 (1983), pp. 917–927.

SIAM J. MATH. ANAL.
Vol. 17, No. 6, November 1986

004

GLOBAL BIFURCATION OF POSITIVE SOLUTIONS IN SOME SYSTEMS OF ELLIPTIC EQUATIONS*

J. BLAT† AND K. J. BROWN†

Abstract. In this paper the structure of the nonnegative steady-state solutions of a system of reaction-diffusion equations arising in ecology is investigated. The equations model a situation in which a predator species and a prey species inhabit the same region and the interaction terms are of Holling–Tanner type so that the predator has finite appetite. Prey and predator birth-rates are treated as bifurcation parameters and the theorems of global bifurcation theory are adapted so that they apply easily to the system. Thus ranges of parameters are found for which there exist nontrivial steady-state solutions.

Key words. reaction-diffusion equations, multiple steady states, global bifurcation, predator-prey

AMS(MOS) subject classifications. Primary 35J65, 32B32, 92A17

1. Introduction. In this paper we study the nonnegative steady-state solutions of the reaction-diffusion system

$$u_t(x,t)-d_1\Delta u(x,t)=au-a_1u^2-a_2uv/(1+mu), \tag{1.1}$$

$$v_t(x,t)-d_2\Delta v(x,t)=bv-b_1v^2+b_2uv/(1+mu)$$

for $x\in D$ and $t\geqq 0$ where D is a bounded region in $\mathbb{R}^n$ $(n=1,2,3)$ with smooth boundary together with boundary conditions

$$u(x,t)=0=v(x,t) \quad \text{for all } x\in\partial D \text{ and } t\geqq 0. \tag{1.2}$$

Equation (1.1) models a situation in which a prey and a predator species with population densities $u(x,t)$ and $v(x,t)$ respectively inhabit the region D. It is assumed that both species diffuse, i.e., move from points of high to points of low population density; the Laplacian terms in (1.1) correspond to this diffusion, the constants d_1 and d_2 giving the rates at which the species diffuse. It is also assumed that in the absence of other species and of diffusion that both species would grow logistically. Thus, in the absence of other factors, the rate of increase of the prey population is given by $au-a_1u^2$. If u is small, this increase is approximately equal to au and the constant a is termed the birth rate of the prey. Because of limited natural resources, the prey population will decrease in size if it becomes too large; we assume throughout that the constant $a_1>0$ so that $au-a_1u^2<0$ for sufficiently large u. Similarly the constant b is termed the birth rate of v and we assume that the constant $b_1>0$. The term $a_2uv/(1+mu)$ represents the rate at which the prey is consumed by the predator and is usually referred to as the Holling–Tanner interaction term; as is reasonable this term increases as either u or v increases. In the classical equations of ecology the corresponding term is simply a_2uv. This classical interaction term has the defect that for a fixed predator population $\lim_{u\to\infty}a_2uv=\infty$ which implies that predators must be capable of consuming prey at an infinitely great rate. For the Holling–Tanner interaction term, however, $\lim_{u\to\infty}a_2uv/(1+mu)=a_2v/m$ and this difficulty does not arise. We assume throughout that the constants $a_2,b_2>0$.

*Received by the editors January 2, 1985.

† Department of Mathematics, Heriot-Watt University, Riccarton, Currie, Midlothian, Scotland, EH14 4AS.

In [2] we used a decoupling technique to study the steady-state solutions of the classical equations of ecology, viz.

$$-d_1\Delta u = au - a_1u^2 - a_2uv, \qquad -d_2\Delta v = bv - b_1v^2 + b_2uv \tag{1.3}$$

with Dirichlet boundary conditions. The general idea is to regard v as a fixed function in the first equation in (1.3) and solve for u, denoting the solution by $u(v)$, i.e., $u(v)$ is the solution of

$$-d_1\Delta u + a_2vu = au - a_1u^2 \quad \text{on } D, \qquad u|_{\partial D} = 0. \tag{1.4}$$

The solutions of (1.4) are easy to describe and a suitable $u(v)$ can be defined. Then $u(v)$ is substituted into the second equation in (1.3) to give a single equation for v. Treating b as a bifurcation parameter and using the results of global bifurcation theory, fairly detailed results are obtained about the solutions of (1.3).

If we fix v in the first equation with the Holling–Tanner interaction term, it seems considerably harder to analyze the solutions of the corresponding equation for u, viz.,

$$-d_1\Delta u = au - a_1u^2 - \frac{a_2uv}{1+mu} \quad \text{on } D, \quad u|_{\partial D} = 0$$

and we are no longer able to use our decoupling technique. In this paper we study nontrivial steady-state solutions of a simplified version of (1.1), i.e.,

$$\begin{aligned} -\Delta u &= au - a_1u^2 - a_2uv/(1+mu) \quad \text{on } D, \qquad u|_{\partial D} = 0, \\ -\Delta v &= bv - b_1v^2 + b_2uv/(1+mu) \quad \text{on } D, \qquad v|_{\partial D} = 0 \end{aligned} \tag{1.5}$$

by applying bifurcation theory directly to the system (1.5). Note that we assume $d_1 = d_2 = 1$; this is done simply for notational convenience and it is straightforward to adapt our proofs to deal also with the general case of unequal diffusion coefficients. If we assume that all the other constants in (1.5) are fixed and treat a as a bifurcation parameter, we can show that (1.5) is equivalent to an operator equation of the form $w - T(a,w) = 0$ where $w = (u,v)$. In the existing literature global bifurcation theorems of the type we require seem only to apply in the cases where the Fréchet derivative $T_w(a,0) = aAw$ (see Rabinowitz [11]) or $T_w(a,0) = aA_1w + A_2w$ (see Chow and Hale [4]) where A, A_1 and A_2 are linear operators. In some situations which we encounter $T_w(a,0)$ depends on a in a more complicated way and in §3 we give a formulation of some standard theorems on bifurcation which can be applied easily to all the cases in which we are interested. In §2 we discuss results we shall require later on linear problems and on the trivial solutions of (1.5). In §§4 and 5 we treat b and a respectively as bifurcation parameters.

We work throughout with Dirichlet boundary conditions. Our results also apply to the more ecologically reasonable cases of Neumann and Robin boundary conditions. However Dirichlet boundary conditions present the hardest mathematical problem and so we concentrate our attention on these. In the case of Neumann boundary conditions all the steady-state solutions we obtain are spatially homogeneous.

A number of other studies have been made on the existence of steady-state solutions of the classical equations of ecology. In Dancer [7] index theory is used to give necessary and sufficient conditions for the existence of nontrivial solutions. Leung [9] has obtained existence and uniqueness results by using iteration methods. Local bifurcation methods for the classical equations describing competing species are used in Cantrell and Cosner [3].

2. Preliminaries. It is well known that the linear eigenvalue problem

$$-\Delta\phi=\lambda\phi \quad \text{on } D, \qquad \phi=0 \quad \text{on } \partial D$$

has an infinite sequence of eigenvalues $\{\lambda_n\}$ such that $0<\lambda_1<\lambda_2\leqq\lambda_3\cdots$ with corresponding eigenfunctions $\phi_1,\phi_2,\phi_3,\cdots$ where $\phi_1(x)>0$ for $x\in D$. Suppose that $q: D\to\mathbb{R}$ is smooth. Then the linear eigenvalue problem

$$-\Delta u+qu=\lambda u \quad \text{on } D, \qquad u=0 \quad \text{on } \partial D \tag{2.1}$$

also has an infinite sequence of eigenvalues which are bounded below. We denote the ith eigenvalue of (2.1) by $\lambda_i(q)$. It is known that $\lambda_1(q)$ is a simple eigenvalue and that the corresponding eigenfunctions do not change sign on D. Clearly $\lambda_1(0)=\lambda_1$ and $\lambda_1(q)$ is an increasing function of q.

If 0 is not an eigenvalue of (2.1), then we can define a corresponding solution operation K, i.e., Kf is the unique solution of

$$-\Delta u+qu=f \quad \text{on } D, \qquad u=0 \quad \text{on } \partial D$$

i.e., K is the inverse of the differential operator $L=-\Delta+q$ associated with Dirichlet boundary conditions. It is well known (see e.g. Amann [1] and Sattinger [12]) that $K: C^1(D)\to C^1(D)$ and $K: L_2(D)\to L_2(D)$ is a compact operator and that $K: C^\alpha(D)\to C_0^{2+\alpha}(D)$ is an isomorphism ($C_0^{2+\alpha}(D)=\{u\in C^{2+\alpha}(D): u(x)=0 \text{ for } x\in\partial D\}$).

Consider now the nonlinear boundary value problem

$$-\Delta u+qu=au-a_1u^2 \quad \text{on } D, \qquad u=0 \quad \text{on } \partial D \tag{2.2}$$

where q is as above and a and a_1 are real numbers with $a_1>0$. It is known that if $a\leqq\lambda_1(q)$ then $u=0$ is the only nonnegative solution of (2.2) whereas if $a>\lambda_1(q)$ then (2.2) has a solution u which is positive on D. Since $u\to(au-a_1u^2)/u$ is a decreasing function, it follows that (see Cohen and Laetsch [5]) for each fixed $a>\lambda_1(q)$ there is a unique solution of (2.2) which is positive on D.

Suppose now that $q=0$. We denote the unique positive solution of

$$-\Delta u=au-a_1u^2 \quad \text{on } D, \qquad u|_{\partial D}=0 \tag{2.3}$$

where $a>\lambda_1$ by u_a. In the (a,u) plane, i.e., $\mathbb{R}\times C^1(D)$, the curve of solutions $a\to u_a$ bifurcates from the zero solution when $a=\lambda_1$. The linearized operator corresponding to u_a is the differential operator L where

$$Lu(x)=-\Delta u(x)-au(x)+2a_1u_a(x)u(x)$$

associated with Dirichlet boundary conditions.

LEMMA 2.1. *All eigenvalues of L are strictly positive.*

Proof. Since

$$-\Delta u_a+(a_1u_a-a)u_a=0 \quad \text{on } D, \qquad u_a|_{\partial D}=0, \tag{2.4}$$

u_a is a positive eigenfunction of $-\Delta+(a_1u_a-a)$ corresponding to the eigenvalue 0 and so $\lambda_1(a_1u_a-a)=0$. Hence $\lambda_1(2a_1u_a-a)>0$ and the result is proved.

We now show that u_a depends continuously on a. Define $F: (\lambda_1,\infty)\times C_0^{2+\alpha}(D)\to C^\alpha(D)$ by

$$F(s,u)=-\Delta u-su+a_1u^2.$$

Clearly F is a C^1 function and $F(s,u_s)=0$ for all $s>\lambda_1$. Choose and fix $a>\lambda_1$. If we denote the Fréchet derivative $F_u(a,u_a)$ by L, then $L: C_0^{2+\alpha}(D)\to C^{\alpha}(D)$ such that $Lu=-\Delta u-au+2a_1u_au$. By Lemma 2.1 L is an isomorphism and so by the Implicit Function Theorem there exists a C^1 function $\phi:\mathbb{R}\to C_0^{2+\alpha}(D)$ defined on a neighbourhood of a such that $\phi(a)=u_a$ and $F(s,\phi(s))=0$. However, $F(s,u)=0$ has the unique solution (s,u_s) close to (a,u_a) and so $\phi(s)=u_s$. This $s\to u_s$ is a C^1 map from $\mathbb{R}$ to $C_0^{2+\alpha}(D)$.

Let $\eta_a=du_a/da$. Differentiation of (2.4) with respect to a and interchange of order of the smooth derivatives involved show that

$$-\Delta\eta_a+(2a_1u_a-a)\eta_a=u_a>0 \quad \text{on } D,$$

i.e.

$$L\eta_a>0 \quad \text{on } D, \qquad \eta_a|_{\partial D}=0.$$

Since by Lemma 2.1 the principal eigenvalue of L on D is positive, there exists a region $\hat{D}$ containing $\overline{D}$ such that the principal eigenvalue of L with Dirichlet boundary conditions on $\hat{D}$ is positive; the corresponding principal eigenfunction ϕ is such that ϕ and $L\phi$ are strictly positive on $\overline{D}$. Hence it follows from the generalized maximum principle (see [10, Chap. 2, Thm. 10]) that η_a/ϕ does not have a nonpositive minimum in D and so $\eta_a>0$. Thus we have

LEMMA 2.2. *The map $a\to u_a$ is a C^1 map from (λ_1,∞) to $C_0^{2+\alpha}(D)$ and, if $\eta_a=du_a/da$, then $\eta_a(x)>0$ for all x in D.*

In a similar way we can define and establish the corresponding properties of v_b, the unique positive solution of

$$-\Delta v=bv-b_1v^2 \quad \text{on } D, \qquad v|_{\partial D}=0$$

when $b>\lambda_1$.

We now discuss the trivial solutions of (1.5). Clearly for all values of a and b there is the zero solution i.e. $u=0$ and $v=0$. When $a>\lambda_1$, there is the semi-trivial solution $u=u_a$ and $v=0$ and, when $b>\lambda_1$, there is the semi-trivial solution $u=0$ and $v=v_b$. We prove the existence of nontrivial solutions by studying the bifurcations which occur from branches of semi-trivial solutions. For this purpose it is necessary to obtain some a priori information about solutions of (1.5).

LEMMA 2.3. *If $a>\lambda_1$, then $(a-\lambda_1)\phi/a_1\leqq u_a\leqq a/a_1$ where ϕ is the principal eigenfunction of $-\Delta$ such that* $\max\phi=1$.

Proof. It is easy to check that $(a-\lambda_1)\phi/a_1$ and a/a_1 are sub and supersolutions of (2.3). But u_a is the unique positive solution of (2.3) and so must lie between the sub- and supersolution.

LEMMA 2.4. *If (u,v) is a nonnegative solution of* (1.5) *such that u is not identically zero, then $a>\lambda_1$.*

Proof. Since u satisfies the first equation in (1.5), it follows that $-\Delta u<au$ on D. Multiplying by u and integrating over D shows that $\int_D|\nabla u|^2dx<a\int_D u^2dx$. But by Poincaré's Inequality $\int_D|\nabla u|^2dx\geqq\lambda_1\int_D u^2dx$ and so $a>\lambda_1$.

The above lemma shows that the prey cannot coexist with the predator if its birth rate is too low. The next lemma gives a priori bounds on the population densities in terms of the birth rates.

LEMMA 2.5. *Suppose (u,v) is a nonnegative solution of* (1.5) *such that $u\not\equiv 0$ and $v\not\equiv 0$. Then*

(i) *$u\leqq u_a$ and $v\leqq b_1^{-1}[b+b_2a/(a_1+ma)]$;*

(ii) *if $b>\lambda_1$, then $v\geqq v_b$.*

Proof. (i) Since $u \not\equiv 0$, it follows that $a > \lambda_1$ and so (2.3) has the unique positive solution u_a. Clearly u is a subsolution of (2.3) and there exist arbitrarily large supersolutions of (2.3). Hence u_a must be greater than or equal to the subsolution u.

As v satisfies the equation

$$-\Delta v = [b - b_1 v + b_2 u/(1+mu)]v \quad \text{on } D \tag{2.5}$$

and $u \leqq u_a \leqq a/a_1$, we have that

$$-\Delta v \leqq [b - b_1 v + ab_2/(a_1 + ma)]v$$

and so $\Delta v > 0$ whenever $v > b_1^{-1}[b + b_2 a/(a_1 + ma)]$ $(= B$ say.) Thus it is impossible that v has a local maximum at x_0 where $v(x_0) > B$ and so we obtain the required upper bound for v.

(ii) Regarding u as a fixed function, v is the unique positive solution of (2.5) with Dirichlet boundary conditions. Clearly v_b is a subsolution of (2.5) and, as there are arbitrarily large constant supersolutions of (2.5), it follows that $v \geqq v_b$.

Finally in this section we make a preliminary investigation of bifurcation from the branch of trivial solutions of the form $u = u_a$, $v = 0$. Writing $u = u_a - U$ and $v = V$, it is easy to check that (u, v) is a nonnegative solution of (1.5) if and only if $0 \leqq U \leqq u_a$, $V \geqq 0$ and (U, V) satisfies

$$\begin{aligned} -\Delta U &= aU - 2a_1 u_a U + a_2 u_a V/(1+mu_a) + f(a, x, U, V), \\ -\Delta V &= bV + b_2 u_a V/(1+mu_a) + g(a, x, U, V) \end{aligned} \tag{2.6}$$

where f and g are smooth functions on $[\lambda_1, \infty) \times D \times \mathbb{R} \times \mathbb{R}$ such that

$$\begin{aligned} f(a, x, U, V) &= a_1 U^2 + a_2[(u_a - U)V/(1 + m(u_a - U)) - u_a V/(1+mu_a)], \\ g(a, x, U, V) &= -b_1 V^2 + b_2[(u_a - U)V/(1 + m(u_a - U)) - u_a V/(1+mu_a)] \end{aligned}$$

for $U \leqq u_a(x) + \frac{1}{2}m^{-1}$. Let $F: (\lambda_1, \infty) \times C^1(D) \times C^1(D) \to C^1(D)$ be defined by

$$[F(a, U, V)](x) = f(a, x, U(x), V(x))$$

and let G be the similar operator corresponding to g. Clearly F and G are continuous and the Fréchet derivatives $F_{(U,V)}(a, 0, 0)$ and $G_{(U,V)}(a, 0, 0)$ are zero. Then equation (2.6) can be written as

$$\begin{aligned} U &= aKU - 2a_1 K(u_a U) + a_2 K[u_a V/(1+mu_a)] + KF(a, U, V), \\ V &= bKV + b_2 K[u_a V/(1+mu_a)] + KG(a, U, V) \end{aligned} \tag{2.7}$$

where K is the inverse of $-\Delta$ with Dirichlet boundary conditions. Clearly (2.7) has the solution $U = 0$, $V = 0$ for all values of a and b and to find possible bifurcation points from the branch of zero solutions it is necessary to investigate the linearisation

$$\begin{aligned} U &= aKU - 2a_1 K(u_a U) + a_2 K[u_a V/(1+mu_a)], \\ V &= bKV + b_2 K[u_a V/(1+mu_a)]. \end{aligned} \tag{2.8}$$

3. Bifurcation theory. Since the dependence of the linearisation (2.8) on a and b is quite complicated, it is necessary to reformulate some of the standard theorems of bifurcation before we can apply them to our equations.

Let X be a Banach space and let $T:\mathbb{R}\times X\to X$ be a compact, continuously differentiable operator such that $T(a,0)=0$. Suppose we can write T as

$$T(a,u)=K(a)u+R(a,u) \tag{3.1}$$

where $K(a)$ is a linear compact operator and the Fréchet derivative $R_u(a,0)=0$. We investigate bifurcation phenomena for the equation

$$u=T(a,u) \tag{3.2}$$

treating a as a bifurcation parameter.

THEOREM 3.1. *Suppose* $(a,0)$ *is a bifurcation point of* (3.2). *Then*

(i) $K(a)$ *has eigenvalue* 1;

(ii) *if* $\{(a_n,u_n)\}$ *is a sequence of nontrivial solutions such that* $a_n\to a$ *and* $u_n\to 0$, *then there exists a subsequence of* $\{u_n\}$, *again denoted by* $\{u_n\}$, *such that* $u_n/\|u_n\|\to u_0$ *where* u_0 *is an eigenvector of* $K(a)$ *corresponding to the eigenvalue* 1.

Proof. Since $u_n-T(a_n,u_n)=0$ for all n, we have that

$$u_n-T_u(a,0)u_n=\int_0^1\left[T_u(a_n,su_n)u_n-T_u(a,0)u_n\right]ds.$$

Let $v_n=u_n/\|u_n\|$. Then

$$v_n-K(a)v_n=\int_0^1\left[T_u(a_n,su_n)v_n-T_u(a,0)v_n\right]ds.$$

As $n\to\infty$, the integral term $\to 0$ and so $\lim_{n\to\infty}(v_n-K(a)v_n)=0$. Since $\{v_n\}$ is bounded and $K(a)$ is compact, there exists a subsequence such that $\{K(a)v_n\}$ is convergent. Hence there exists a subsequence of $\{v_n\}$ converging to u_0 say. As $\|v_n\|=1$ for all n, $u_0\neq 0$ and clearly $u_0-K(a)u_0=0$.

We now give a sufficient condition for bifurcation to occur and at the same time obtain a global bifurcation result. First we recall the notions of multiplicity of an eigenvalue and the index of a fixed point.

Let $K:X\to X$ be a compact linear operator and let λ_0 be a nonzero eigenvalue of K. Then the null space of $K-\lambda_0 I$ denoted by $N(K-\lambda_0 I)$ is nonempty. We define the generalized null space of λ_0 as $M(K;\lambda_0)=\bigcup_{p=1}^{\infty}N(K-\lambda_0 I)^p$. It is well known that

$$N(K-\lambda_0 I)\subseteq N(K-\lambda_0 I)^2\subseteq N(K-\lambda_0 I)^3\subseteq\cdots$$

and that there exists P such that $N(K-\lambda_0 I)^p\subsetneq N(K-\lambda_0 I)^{p+1}$ for $p<P$ but $N(K-\lambda_0 I)^P=N(K-\lambda_0 I)^{P+1}=N(K-\lambda_0 I)^{P+2}=\cdots$ and so

$$M(K;\lambda_0)=\bigcup_{p=1}^{P}N(K-\lambda_0 I)^p.$$

Thus $\dim M(K;\lambda_0)<\infty$ and we define the algebraic multiplicity of λ_0 as equal to $\dim M(K;\lambda_0)$. If λ_0 has algebraic multiplicity 1, we say that λ_0 is a simple eigenvalue. Clearly λ_0 is a simple eigenvalue if and only if $\dim N(K-\lambda_0 I)=\dim N(K-\lambda_0 I)^2=1$. It is easy to show that $N(K-\lambda_0 I)=N(K-\lambda_0 I)^2$ if and only if $N(K-\lambda_0 I)\cap R(K-\lambda_0 I)=\{0\}$ where $R(K-\lambda_0 I)$ denotes the range space of $K-\lambda_0 I$. Thus λ_0 is a simple eigenvalue of K if and only if $\dim N(K-\lambda_0 I)=1$ and $N(K-\lambda_0 I)\cap R(K-\lambda_0 I)=\{0\}$.

Suppose now that $I-K: X \to X$ is a bijection. Then it is well known that the Leray Schauder degree $\deg(I-K,B,0)=(-1)^p$ where B is a ball centre 0 in X and $p=$ sum of the algebraic multiplicities of the eigenvalues of K which are >1 (see Krasnoselskii [8]). Suppose $T: X \to X$ is a compact differentiable operator. If x_0 is an isolated fixed point of T, we define the index of T at x_0 as $i(T,x_0)=\deg(I-T,B,x_0)$ where B is a ball centre x_0 such that x_0 is the only fixed point of T in B. If x_0 is a fixed point of T such that $I-T'(x_0)$ is invertible, then x_0 is an isolated fixed point of T and

$$i(T,x_0)=\deg(I-T,B,x_0)=\deg\big(I-T'(x_0),\hat{B},0\big)$$

where B is a sufficiently small ball centre x_0 and $\hat{B}$ is a ball centre 0.

We now state the result we shall use on global bifurcation. Suppose $T: \mathbb{R}\times X \to X$ is as given by (3.1).

THEOREM 3.2. *Let a_0 be such that $I-K(a)$ is invertible if $0<|a-a_0|<\varepsilon$ for some $\varepsilon>0$. Suppose $i(T(a,\cdot),0)$ is constant on $(a_0-\varepsilon,a_0)$ and on $(a_0,a_0+\varepsilon)$ such that, if $a_0-\varepsilon<a_1<a_0<a_2<a_0+\varepsilon$, then $i(T(a_1,\cdot),0)\neq i(T(a_2,\cdot),0)$. Then there exists a continuum C in the $(a-u)$-plane of solutions of* (3.2) *such that one of the following alternatives holds*

(i) *C joins $(a_0,0)$ to $(\hat{a},0)$ where $I-K(\hat{a})$ is not invertible.*

(ii) *C joins $(a_0,0)$ to ∞ in $\mathbb{R}\times X$.*

The above results can be proved by using exactly the same argument as in Rabinowitz [11]. The index $i(T(a,\cdot),0)$ can be calculated by investigating the eigenvalues of $K(a)$.

4. Structure of solutions with b as bifurcation parameter. In this section we shall regard b as a bifurcation parameter and suppose that all other constants are fixed. For all values of b we have the branch of zero solutions of (1.5) $S_0=\{(b,0,0): b\in\mathbb{R}\}$. When b crosses λ_1, there bifurcates from S_0 the branch of semi-trivial solutions $S_1=\{(b,0,v_b): b>\lambda_1\}$. Lemma 2.4 shows that, when a is fixed $\leqq\lambda_1$, then all nonnegative solutions of (1.5) lie on either S_0 or S_1. For the rest of this section we suppose that a is fixed $>\lambda_1$ so that we also have the branch of semi-trivial solutions $S_2=\{(b,u_a,0): b\in\mathbb{R}\}$. We show that there is a continuum of nontrivial solutions (i.e. in which neither u nor v is identically zero) joining S_1 and S_2.

First we use the result of Crandall and Rabinowitz [6] on bifurcation from a simple eigenvalue to obtain a local result on bifurcation from S_2. Motivated by (2.7), we define $T: \mathbb{R}\times C^1(D)\times C^1(D)\to C^1(D)\times C^1(D)$ by $T(b,u,v)=(aKu-2a_1K(u_au)+a_2K[u_av/(1+mu_a)]+KF(a,u,v),\ bKv+b_2K[u_av/(1+mu_a)]+KG(a,u,v))$ where K is the inverse of $-\Delta$ with Dirichlet boundary conditions. Let $H=I-T$. Then $H(b,u,v)=0$ with $0\leqq u\leqq u_a$ and $v\geqq 0$ if and only if (b,u_a-u,v) is a nonnegative solution of (1.5). It is easy to see that H is a C^1 function with $H(b,0,0)=0$. Straightforward computations show that H has Fréchet derivative

(4.1)

$$H_{(u,v)}(b,0,0)(\phi,\psi)$$
$$=\big(\phi-aK\phi+2a_1K(u_a\phi)-a_2K[u_a\psi/(1+mu_a)],\ \psi-bK\psi-b_2K[u_a\psi/(1+mu_a)]\big).$$

For notational convenience let $q(x)=u_a(x)/(1+mu_a(x))$. Let $b_0=\lambda_1(-b_2q)$ and let ψ_1 denote a corresponding nonnegative principal eigenfunction corresponding to principal eigenvalue b_0 of

$$-\Delta\psi-b_2q\psi=b\psi \quad \text{on } D, \qquad \psi|_{\partial D}=0.$$

It is easy to check that $N(H_{(u,v)}(b_0,0,0))=\operatorname{span}\{(\phi_1,\psi_1)\}$ where $\phi_1=a_2K_1[q\psi_1]$ and K_1 denotes the inverse of $-\Delta-a+2a_1u_a$ with Dirichlet boundary conditions. Thus $\dim N(H_{(u,v)}(b_0,0,0))=1$. It follows from the properties of compact operators that the codimension of $R(H_{(u,v)}(b_0,0,0))=1$.

Using (4.1), it is easy to see that the Fréchet derivative

$$H_{b,(u,v)}(b_0,0,0)(\phi_1,\psi_1)=(0,-K\psi_1).$$

Suppose that $(0,-K\psi_1)\in R(H_{(u,v)}(b_0,0,0))$. Then there exists $\psi\in C^1(D)$ such that

$$\psi-bK\psi-b_2K[q\psi]=-K\psi_1$$

and so

$$-\Delta\psi-b\psi-b_2q\psi=-\psi_1. \tag{4.2}$$

Multiplying (4.2) by ψ_1 and integrating over D shows that $\int_D\psi_1^2\,dx=0$ and this is impossible. Thus $(0,-K\psi_1)\notin R(H_{(u,v)}(b_0,0,0))$.

We have shown that H satisfies all the hypotheses of Theorem 1.7 of [6]. Thus there exists a real interval $(-\varepsilon,\varepsilon)$ and functions $b:(-\varepsilon,\varepsilon)\to\mathbb{R}$, $u,v:(-\varepsilon,\varepsilon)\to C^1(D)$ such that the nontrivial zeros of H close to $(b_0,0,0)$ lie on the curve $\{(b(s),s\phi_1+su(s),s\psi_1+sv(s)):-\varepsilon<s<\varepsilon\}$ where $b(0)=b_0$, $u(0)=v(0)=0$. It follows that for the system of equations (1.5) bifurcation occurs from the branch of semi-trivial solutions S_2 at $(b_0,u_a,0)$ and close to the bifurcation point the nontrivial solutions lie on the curve $\{(b(s),u_a-s\phi_1-su(s),s\psi_1+sv(s)):-\varepsilon<s<\varepsilon\}$. Points on the curve with $s>0$ correspond to nontrivial, nonnegative solutions of (1.5).

We now investigate the global nature of the above curve of nontrivial, nonnegative solutions in the $b-(u,v)$ plane, i.e., in $\mathbb{R}\times C^1(D)\times C^1(D)$. Theorems 4.1 and 4.2 give limitations on the values of b for which such solutions can exist.

THEOREM 4.1. *If (b,u,v) be a nontrivial, nonnegative solution of* (1.5), *then $b>b_0$ (i.e. $b>\lambda_1(-b_2q)$).*

Proof. We have that

$$-\Delta v-\left[b_2u/(1+mu)\right]v=bv-b_1v^2.$$

Since $u\leqq u_a$, it follows that

$$-\Delta v-b_2qv\leqq bv-b_1v^2.$$

Multiplying by v and integrating over D, we obtain

$$-\int_D(\Delta v+b_2qv)v\,dx<b\int_D v^2\,dx.$$

By the spectral theorem

$$-\int_D(\Delta v+b_2qv)v\,dx\geqq b_0\int_D v^2\,dx$$

and so it follows that $b>b_0$.

The above result shows that the bifurcation of nonnegative solutions from S_2 at $(b_0,u_a,0)$ must be to the right. The next result shows that the branch of nontrivial solutions cannot extend too far to the right.

THEOREM 4.2. *There exists $M>0$ such that, if (b,u,v) is a nontrivial, nonnegative solution of* (1.5), *then $b\leqq M$.*

Proof. By Lemma 2.3 $v_b \geqq (b-\lambda_1)\phi/b_1$ where ϕ is the principal eigenfunction of $-\Delta$ with $\sup\phi = 1$. If α is a positive constant, it follows that $\lim_{b\to\infty}\lambda_1(\alpha v_b) = \infty$ (see [2, Thm. 3.4]). Choose M such that $\lambda_1(a_1a_2v_b/(a_1+ma)) > a$ whenever $b > M$.

Suppose (b,u,v) is a nontrivial, nonnegative solution of (1.5). Then u is a nontrivial solution of

$$-\Delta u + [a_2v/(1+mu)]u = au - a_1u^2$$

and so $a > \lambda_1(a_2v/(1+mu))$. Since $v \leqq v_b$ and $u \leqq a/a_1$, it follows that $a > \lambda_1(a_1a_2v_b/(a_1+ma))$. Hence we must have that $b < M$.

We now show that T satisfies the hypotheses of our global bifurcation result, i.e., Theorem 3.2. We can write T as

$$T(b,u,v) = K(b)(u,v) + R(b,u,v)$$

where $K(b)$ is the compact linear operator such that $K(b)(u,v) = (aKu - 2a_1K(u_au) + a_2K(qv),\ bKv + b_2K(qv))$ and $K(b)$ and $R(b,u,v)$ satisfy the conditions given at the start of §3.1. In order to show that the hypotheses of Theorem 3.2 are satisfied, we must calculate the index $i(T(b,\cdot),0)$ when b is close to b_0. This index is equal to $(-1)^\beta$ where β is the sum of the algebraic multiplicities of eigenvalues of $K(b) > 1$.

Suppose that $\mu > 0$ is an eigenvalue of $K(b)$. Then there exists a nonzero function v such that

$$bKv + b_2K(qv) = \mu v$$

and so

$$-\mu\Delta v - b_2qv = bv \quad \text{on } D, \qquad v = 0 \quad \text{on } \partial D, \tag{4.3}$$

i.e. b is an eigenvalue of (4.3). Conversely, if $\mu \geqq 1$ and b is an eigenvalue of (4.3) with corresponding eigenfunction v, then (u,v) is an eigenfunction of $K(b)$ corresponding to the eigenvalue μ where u is the unique solution of

$$-\mu\Delta u - au + 2a_1u_au = a_2qv \quad \text{on } D, \qquad u = 0 \quad \text{on } \partial D.$$

Note that, since all eigenvalues of $-\Delta - a + 2a_1u_a$ are positive by Lemma 2.1 and $\mu > 1$, it follows that $-\mu\Delta - a + 2a_1u_a$ is invertible. The eigenvalues of (4.3) form an increasing sequence $\gamma_1(\mu) < \gamma_2(\mu) \leqq \gamma_3(\mu) \leqq \cdots$ and the variational characterisation of eigenvalues shows that $\mu \to \gamma_i(\mu)$ is a continuous increasing function. Thus $\mu \geqq 1$ is an eigenvalue of $K(b)$ if and only if $b = \gamma_i(\mu)$ for some μ. Clearly $\gamma_i(1) = \lambda_i(-b_2q)$.

Suppose that $b < b_0$, i.e., $b < \lambda_1(-b_2q)$. Hence $b < \gamma_1(1)$ and so $b < \gamma_i(\mu)$ for $i = 1, 2, \cdots$ and $\mu \geqq 1$. Hence $K(b)$ has no eigenvalues > 1 and so $i(T(b,\cdot),0) = 1$.

Suppose $b_0 < b < \lambda_2(-b_2q)$. Then $\gamma_1(1) < b < \gamma_2(1)$. Since $\mu \to \gamma_1(\mu)$ is increasing with $\lim_{\mu\to\infty}\gamma_1(\mu) = \infty$, there exists a unique $\mu > 1$ (μ_1 say) such that $b = \gamma_1(\mu_1)$. Since $b < \gamma_2(1)$, it follows that $b < \gamma_i(\mu)$ for $i = 2, 3, \cdots$ and $\mu \geqq 1$. Thus μ_1 is the only eigenvalue of $K(b)$ which is greater than 1. We now show that μ_1 is a simple eigenvalue of $K(b)$. The discussion above shows that $N(K - \mu_1 I) = \text{span}\{(\phi,\psi)\}$ where ψ is the principal eigenfunction corresponding to the eigenvalue b of

$$-\mu_1\Delta v - b_2qv = bv \quad \text{on } D, \qquad v = 0 \quad \text{on } \partial D$$

and $\phi = a_2K_1[q\psi]$ and K_1 denotes the inverse of $-\mu_1\Delta - a + 2a_1u_a$. Thus $\dim N(K(b) - \mu_1 I) = 1$. Suppose that $(\phi,\psi) \in R(K(b) - \mu_1 I)$. Then there exists v such that

$$bKv + b_2K(qv) - \mu_1 v = \psi.$$

Hence

$$-\mu_1\Delta v - bv - b_2qv = -\Delta\psi = -\mu_1^{-1}(b\psi + b_2q\psi) \quad \text{on } D.$$

Multiplying by ψ and integrating over D shows that $\int_D (b+b_2q)\psi^2\,dx = 0$ which is impossible. Hence $R(K(b)-\mu_1 I)\cap N(K(b)-\mu_1 I)=\{0\}$ and so μ_1 is a simple eigenvalue of $K(b)$. Thus $i(T(b,\cdot),0)=-1$ whenever $b_0 < b < \lambda_2(-b_2q)$.

Therefore Theorem 3.2 can be applied to T. Thus there exists a continuum C_0 of solutions of $(u,v)=T(b,u,v)$ in the $b-(u,v)$ plane, i.e., in $\mathbb{R}\times C^1(D)\times C^1(D)$ emanating from $(b_0,0,0)$ and either joining with $(\hat{b},0,0)$ where $I-K(\hat{b})$ is not invertible or joining with ∞. Close to the bifurcation point all solutions lie on the curve whose existence we proved by using the Crandall and Rabinowitz theorem. Let C_1 be the maximal continuum of solutions contained in $C_0 - \{(b(s), s\phi_1 + su(s), s\psi_1 + sv(s)) : -\varepsilon < s \leqq 0\}$. Then close to the bifurcation point $(b_0,0,0)$ C_1 consists of the curve $\{(b(s), s\phi_1 + su(s), s\psi_1 + sv(s)) : 0 < s < \varepsilon\}$ and it can be shown by a reflection argument exactly as in Rabinowitz [11] that C_1 either satisfies one of the same alternatives as C_0 or contains a pair of points of the form (b,u,v) and $(b,-u,-v)$ where $(u,v)\neq(0,0)$. Let $C=\{(b,u_a-u,v):(b,u,v)\in C_1\}$. Clearly, if $u,v\geqq 0$ and $(b,u,v)\in C$, then (b,u,v) is a solution of system (1.5). Let

$$P_1 = \{u\in C^1(D) : u(x) > 0 \text{ for } x\in D \text{ and } \partial u/\partial n(x) < 0 \text{ for } x\in\partial D\}$$

and let

$$P = \{(b,u,v) : b\in\mathbb{R} \text{ and } u,v\in P_1\}.$$

Clearly $C\subset P$ in a neighbourhood of the bifurcation point $(b_0,u_a,0)$. However

THEOREM 4.3. *$C-\{(b_0,u_a,0)\}$ is not contained in P.*

Proof. Suppose $C-\{(b_0,u_a,0)\}$ is contained in P; we shall obtain a contradiction. By the previous discussion the continuum $C-\{(b_0,u_a,0)\}$ must

(i) contain points of the form (b,u_a-u,v) and $(b,u_a+u,-v)$ or

(ii) join up with a bifurcation point of the form $(\hat{b},u_a,0)$ where $\hat{b}\neq b_0$ and $I-K(\hat{b})$ is not invertible or

(iii) join $(b_0,u_a,0)$ with ∞.

Since the continuum is contained in P neither (i) nor (ii) is possible. By Theorem 4.1 and Theorem 4.2 we must have that $b_0 < b < M$ whenever $(b,u,v)\subset C$. Therefore by Lemma 2.5 there exists a constant $M_1>0$ such that $|u(x)|$, $|v(x)|\leqq M_1$ for all $x\in D$ whenever $(b,u,v)\in C$. It follows from standard bootstrapping arguments that C is bounded in $\mathbb{R}\times C^1(D)\times C^1(D)$ and so (iii) is also impossible. This is a contradiction and so the continuum is not contained in P.

THEOREM 4.4. *C joins with S_1.*

Proof. Since $C-\{(b_0,u_a,0)\}$ is not contained in P, there exists $(\hat{b},\hat{u},\hat{v})\in [C-\{(b_0,u_a,0)\}]\cap\partial P$ which is the limit of a sequence $\{(b_n,u_n,v_n)\}\subseteq C\cap P$. As $(\hat{b},\hat{u},\hat{v})\in\partial P$, either $\hat{u}\in\partial P_1$ or $\hat{v}\in\partial P_1$.

Suppose $\hat{v}\in\partial P_1$. Then $\hat{v}(x)\geqq 0$ for $x\in D$ and either $\hat{v}(x)=0$ for some $x\in D$ or $\partial\hat{v}/\partial n(x)=0$ for some $x\in\partial D$. It follows from the second equation in system (1.5) that

$$-\Delta\hat{v} + \left[M - b + b_1\hat{v} - b_2\hat{u}/(1+m\hat{u})\right]\hat{v} = M\hat{v}\geqq 0 \quad \text{on } D$$

where M is a constant chosen sufficiently large so that the term in the square brackets is positive for all $x\in D$. It follows from the maximum principle that $\hat{v}\equiv 0$. A similar but simpler argument shows that if $\hat{u}\in\partial P_1$ then $\hat{u}\equiv 0$. Thus $\hat{u}\equiv 0$ or $\hat{v}\equiv 0$.

Suppose that $\hat{u} \equiv 0$ and $\hat{v} \equiv 0$. Then $(\hat{b}, \hat{u}, \hat{v})$ lies on the branch of trivial solutions $S_0 = \{(b,0,0) : b \in \mathbb{R}\}$. The only nontrivial, nonnegative solutions which are close to S_0 lie on the semi-trivial branch $S_1 = \{(b,0,v_b) : b \geqq \lambda_1\}$ and so there cannot exist a sequence in P converging to $(\hat{b},0,0)$. Hence either $\hat{u}$ or $\hat{v}$ is nonzero.

Suppose that $\hat{u}$ is nonzero and $\hat{v} \equiv 0$. Then $(\hat{b}, \hat{u}, \hat{v})$ lies on $S_2 = \{(b, u_a, 0) : b \in \mathbb{R}\}$ and so is a bifurcation point on S_2 from which bifurcate nontrivial and nonnegative solutions. Therefore by Theorem 3.1 $\hat{b}$ is such that 1 is an eigenvalue of $K(\hat{b})$ with corresponding eigenfunctions which are nonnegative on D. If (u,v) is a nonnegative eigenfunction corresponding to $\hat{b}$, then v satisfies

$$-\Delta v - b_2 q v = \hat{b} v \quad \text{on } D, \qquad v = 0 \quad \text{on } \partial D$$

and since v is nontrivial and nonnegative, it follows that $\hat{b} = \lambda_1(-b_2 q) = b_0$ and this is impossible.

Thus the only remaining possibility is that $\hat{v}$ is nonzero and $\hat{u} \equiv 0$. In this case $(\hat{b}, \hat{u}, \hat{v})$ must lie on S_1. Hence C joins up with S_1.

It is possible to use our methods to analyze the bifurcation which occurs when C joins up with S_1. The arguments involved are very similar to those we develop in the next section and so we omit the details here. In fact C joins S_1 when b is such that $a = \lambda_1(a_2 v_b)$ (when $b = b_1$ say). The argument in Theorem 4.4 shows that, if $(b,u,v) \in C \cap \partial P$, then $(b,u,v) \in S_1$. Thus C provides a continuum of nontrivial, nonnegative solutions joining $(b_0, u_a, 0)$ on S_2 to the point $(b_1, 0, \hat{v})$ on S_1. In particular we can conclude

THEOREM 4.5. *The system of equations* (1.5) *has a nontrivial, nonnegative solution provided* $b_0 < b < b_1$.

5. Structure of solutions with a as bifurcation parameter. We now treat a as a bifurcation parameter and assume that all the other constants are fixed. The decoupling technique of [2] works in this case. Suppose $b > \lambda_1$. Then (1.5) has a continuum of semi-trivial solutions $S_1 = \{(a,0,v_b) : a \in \mathbb{R}\}$ and it can be proved as in [2] that there is a continuum C of nontrivial, nonnegative solutions bifurcating from S_1 at $(\lambda_1(a_2 v_b), 0, v_b)$ such that C does not join up with any other continuum and goes to ∞ as $a \to \infty$. Thus the following result holds.

THEOREM 5.1. *If* $b > \lambda_1$, *then the system of equations* (1.5) *has a nontrivial, nonnegative solution provided* $a > \lambda_1(a_2 v_b)$.

The above result could also be established by using an argument similar to that of the preceding section.

Suppose now that $b < \lambda_1$. In this case we have the continuum of trivial solutions $S_0 = \{(a,0,0) : a \in \mathbb{R}\}$ and the continuum of semi-trivial solutions $S_1 = \{(a, u_a, 0) : a \geqq \lambda_1\}$. In [2], using decoupling techniques for the classical predator-prey equations it was shown that the stability of the semi-trivial solution $(u_a, 0)$ changes as a is increased and this indicates that a continuum of nontrivial, nonnegative solutions bifurcates from S_1. We now use bifurcation techniques similar to those of the preceding section to make a direct investigation of this continuum for the more complicated system (1.5).

As we are linearizing about the same solution as in the previous section, bifurcation seems likely to occur at values of a such that $b = \lambda_1(-b_2 u_a/(1 + m u_a))$. Modifying our notation slightly from that used in the previous section in order to highlight the dependence on a, we let $q_a(x) = u_a(x)/(1 + m u_a(x))$. Since u_a is an increasing function of a, q_a is also an increasing function of a. Clearly $q_a(x) \leqq m^{-1}$ for all a. Hence $\lambda_1(-b_2 q_a)$ is a decreasing function of a and $\lambda_1(-b_2 q_a) \geqq \lambda_1 - b_2/m$ for all a. Thus

$\lim_{a\to\infty}\lambda_1(-b_2q_a)$ exists and $\lim_{a\to\infty}\lambda_1(-b_2q_a)\geqq\lambda_1-b_2/m$. By Lemma 2.3 u_a converges uniformly to ∞ on any compact subset of D and so b_2q_a converges uniformly to b_2/m. If ϕ denotes the principal eigenfunction of $-\Delta$ with zero boundary conditions such that $\int_D\phi^2\,dx=1$, then $\int_D|\nabla\phi|^2\,dx=\lambda_1$. Since $\lim_{a\to\infty}\int_D b_2q_a\phi^2\,dx=b_2/m\int_D\phi^2\,dx=b_2/m$, it follows that

$$\lim_{a\to\infty}\int_D\left(|\nabla\phi|^2-b_2q_a\phi^2\right)dx=\lambda_1-b_2/m.$$

Hence by the variational characterisation of eigenvalues $\lim_{a\to\infty}\lambda_1(-b_2q_a)\leqq\lambda_1-b_2/m$.

Thus we have shown

LEMMA 5.1. $\lim_{a\to\infty}\lambda_1(-b_2q_a)=\lambda_1-b_2/m$.

Our first theorem shows that if the predator birth rate is too low then no nontrivial solutions exist. This result differs from what occurs in the classical case where, however negative the predator birth rate, nontrivial solutions exist provided the prey birth rate is sufficiently large.

THEOREM 5.2. *If* (a,u,v) *is a nonnegative solution of* (1.5) *with* $v\not\equiv 0$, *then* $b>\lambda_1-b_2/m$.

Proof. Since

$$-\Delta v-\left[b_2u/(1+mu)\right]v=bv-b_1v^2,$$

it follows that

$$-\Delta v-b_2/mv\leqq bv-b_1v^2.$$

Multiplying by v and integrating over D, we obtain

$$-\int_D\Delta v\cdot v\,dx-b_2/m\int_D v^2\,dx<b\int_D v^2\,dx.$$

Since $-\int_D\Delta v\cdot v\,dx\geqq\lambda_1\int_D v^2\,dx$, it follows that $b>\lambda_1-b_2/m$.

The above theorem shows that, if we fix $b\leqq\lambda_1-b_2/m$, there can be no bifurcation of nontrivial solutions from S_1. From now on we suppose that b is fixed such that $b>\lambda_1-b_2/m$. Since $a\to\lambda_1(-b_2q_a)$ is a decreasing function which equals λ_1 when $a=\lambda_1$ and tends to λ_1-b_2/m as $a\to\infty$, there is a unique value of $a>\lambda_1$, say α, such that $b=\lambda_1(-b_2q_\alpha)$. We show that bifurcation from S_1 occurs at $(\alpha,u_\alpha,0)$.

Motivated by equation (2.7) as in the previous section but now interested in varying a rather than b, we define $T:\mathbb{R}\times C^1(D)\times C^1(D)\to C^1(D)\times C^1(D)$ by

$$T(a,u,v)$$
$$=\left(aKu-2a_1K(u_au)+a_2K(q_av)+KF(a,u,v),\ bKv+b_2K(q_av)+KG(a,u,v)\right)$$

and let $H=I-T$. Then $H(a,u,v)=0$ with $0\leqq u\leqq u_a$ and $v\geqq 0$ if and only if (a,u_a-u,v) is a nonnegative solution of (1.5). Clearly $H(a,0,0)=0$ and it follows from Lemma 2.2 that H is a C^1 function. The Fréchet derivative with respect to (u,v) is

$$H_{(u,v)}(a,0,0)(\phi,\psi)=\left(\phi-aK\phi+2a_1K(u_a\phi)-a_2K(q_a\psi),\ \psi-bK\psi-b_2K(q_a\psi)\right).$$

We have that $N(H_{(u,v)}(\alpha,0,0))=\operatorname{span}\{(\phi_1,\psi_1)\}$ where ψ_1 is a nonnegative eigenfunction corresponding to the principal eigenvalue b $(=\lambda_1(-b_2q_\alpha))$ of

$$-\Delta\psi-b_2q_\alpha\psi=b\psi\quad\text{on } D,\qquad \psi|_{\partial D}=0$$

and $\phi_1 = a_2 K_1(q_\alpha \psi_1)$ where K_1 denotes the inverse of $-\Delta - a + 2a_1 u_a$ with Dirichlet boundary conditions. Thus $\dim N(H_{(u,v)}(\alpha,0,0)) = 1$ and it follows that the codimension of $R(H_{(u,v)}(\alpha,0,0)) = 1$.

It is easy to check that further differentiation with respect to a gives

$$H_{a,(u,v)}(\alpha,0,0)(\phi_1,\psi_1)$$

$$= \left(-K\phi_1 + 2a_1 K(u'_\alpha \phi_1) - a_2 K\left[\psi_1 u'_\alpha/(1+mu_\alpha)^2\right],\ -b_2 K\left[\psi_1 u'_\alpha/(1+mu_\alpha)^2\right]\right)$$

where $u'_a = du_a/da$. Suppose that $H_{a,(u,v)}(\alpha,0,0)(\phi_1,\psi_1) \in R(H_{(u,v)}(\alpha,0,0))$. Then there exists v such that

$$v - bKv - b_2 K(q_\alpha v) = -b_2 K\left[\psi_1 u'_\alpha/(1+mu_\alpha)^2\right]$$

and so

$$-\Delta v - bv - b_2 q_a v = -b_2 \psi_1 u'_\alpha/(1+mu_\alpha)^2.$$

Multiplying by ψ_1 and integrating over D shows that

$$0 = -b_2 \int_D u'_\alpha \psi_1^2/(1+mu_\alpha)^2\, dx.$$

But by Lemma 2.2 u'_α is positive on D and so we have a contradiction.

Thus the hypotheses of the Crandall–Rabinowitz theorem are satisfied and so there exists a curve of nontrivial, nonnegative solutions of (1.5) bifurcating from S_1 at $(\alpha, u_\alpha, 0)$. We shall investigate the global nature of this continuum. First we show that bifurcation of nonnegative solutions is to the right.

THEOREM 5.3. *If (a,u,v) is a nontrivial, nonnegative solution of* (1.5), *then $a > \alpha$.*

Proof. Suppose $a \leqq \alpha$. Then $b = \lambda_1(q_\alpha) \leqq \lambda_1(q_a)$. But by Theorem 4.1 system (1.5) has nontrivial, nonnegative solutions only when $b > \lambda_1(q_a)$. Thus, if (a,u,v) is nontrivial, $a > \alpha$.

We now compute $i(T(a,\cdot),0)$ so that we can apply our global bifurcation result. This index is $(-1)^\beta$ where β is the sum of the algebraic multiplicities of the eigenvalues of $K(a) > 1$ where $K(a)$ is the compact linear operator

$$K(a)(u,v) = \left(aKu - 2a_1 K(u_a u) + a_2 K(q_a v),\ bKv + b_2 K(q_a v)\right).$$

If $\mu > 0$ is an eigenvalue of $K(a)$, then b must be an eigenvalue of

$$-\mu\Delta v - b_2 q_a v = \lambda v \quad \text{on } D, \qquad v|_{\partial D} = 0. \tag{5.1}$$

Conversely, if $\mu \geqq 1$ and b is an eigenvalue of (5.1) with corresponding eigenfunction v, then (u,v) is an eigenfunction of $K(a)$ corresponding to the eigenvalue μ where $u = (-\mu\Delta - a + 2a_1 u_a)^{-1} v$, the inverted differential operator corresponding to zero boundary conditions.

Suppose that $a < \alpha$ and $\mu \geqq 1$ is an eigenvalue of $K(a)$. Then b is an eigenvalue of (5.1) and, since $\mu \geqq 1$, it follows that $b >$ the least eigenvalue of $-\Delta - b_2 q_a$, i.e. $b > \lambda_1(-b_2 q_a)$. But $b = \lambda_1(-b_2 q_\alpha) < \lambda_1(-b_2 q_a)$ and this is a contradiction. Hence, if $a < \alpha$, $K(a)$ has no eigenvalues > 1 and so $i(T(a,\cdot),0) = 1$.

Now suppose that a is such that $\lambda_1(-b_2q_a)<b<\lambda_2(-b_2q_a)$, i.e. a lies in an open interval with left-hand end point α. An argument similar to that used in the preceding section shows that $K(a)$ has a unique eigenvalue μ_1 which is greater than 1 and that this eigenvalue is simple. Therefore $i(T(a,\cdot),0)=-1$ and so Theorem 3.2 can be applied to T.

By arguments similar to those used in the preceding section it can be proved that there exists a continuum C in $\mathbb{R}\times C^1(D)\times C^1(D)$ emanating from $(\alpha,u_\alpha,0)$ such that

(i) if $(a,u,v)\in C$, then $(u_a-u,v)=T(a,u_a-u,v)$;

(ii) if $(a,u,v)\in C$ and $u,v\geqq 0$, then (a,u,v) is a solution of (1.5);

(iii) close to the bifurcation point $(\alpha,u_\alpha,0)$, C consists of the points (a,u,v) on the curve given by the Crandall and Rabinowitz theorem with $v\geqq 0$.

We now show that C does not join up with any other continuum but extends to $a=\infty$.

THEOREM 5.4. (i) *If* $(a,u,v)\in C-\{(\alpha,u_\alpha,0)\}$, *then* $u,v\in P_1$, *i.e.* $u,v>0$ *on* D *and* $\partial u/\partial n$, $\partial v/\partial n<0$ *on* ∂D.

(ii) $\{a:(a,u,v)\in C\}=[\alpha,\infty)$.

Proof. (i) Suppose that C contains a point $(a,u,v)\neq(\alpha,u_\alpha,0)$ which lies outside of P. Then there exists a point $(\hat a,\hat u,\hat v)\in C-\{(\alpha,u_\alpha,0)\}\cap\partial P$ which is the limit of a sequence of points $\{(a_n,u_n,v_n)\}$ in $C\cap P$. It follows as in the previous section that $\hat u\equiv 0$ or $\hat v\equiv 0$.

Suppose that $\hat u\equiv 0$ and $\hat v\equiv 0$. Then $(\hat a,\hat u,\hat v)=(\hat a,0,0)$ and so $(\hat a,\hat u,\hat v)$ lies on the trivial branch of solutions S_0. The only nontrivial, nonnegative solutions which are close to S_0 lie on the semi-trivial branch $S_1=\{(a,u_a,0):a\geqq\lambda_1\}$ and so there cannot exist a sequence in $C\cap P$ converging to $(\hat a,\hat u,\hat v)$. Therefore it is impossible that both $\hat u$ and $\hat v$ are identically zero.

Suppose that $\hat u\equiv 0$. Then

$$-\Delta\hat v=b\hat v-b_1\hat v^2 \quad \text{on } D, \qquad \hat v|_{\partial D}=0$$

and so, since $b<\lambda_1$, $\hat v\equiv 0$. Therefore $\hat u$ is not identically zero.

Suppose that $\hat v\equiv 0$. Then $(\hat a,\hat u,\hat v)\in S_1$ and there bifurcate from $(\hat a,\hat u,\hat v)$ nontrivial, nonnegative solutions. Therefore by Theorem 3.1 $\hat a$ is such that 1 is an eigenvalue of $K(\hat a)$ with corresponding eigenfunctions which are nonnegative on D. Thus $\hat a$ must be such that $b=\lambda_1(-b_2q_a)$ when $a=\hat a$. Hence $\hat a=\alpha$ and $(\hat a,\hat u,\hat v)=(\alpha,u_\alpha,0)$ which is impossible.

Therefore, if $(a,u,v)\in C-\{(\alpha,u_\alpha,0)\}$, then $(a,u,v)\in P$.

(ii) C must satisfy one of the three alternatives discussed in the preceding section. Because of (i) above, C contains no pairs of points of the form (a,u_a-u,v) and $(a,u_a+u,-v)$ and C cannot join up with another bifurcation point of the form $(a,u_a,0)$ on S_1. Hence C joins $(\alpha,u_\alpha,0)$ to ∞. By Theorem 5.3 we have that $a\geqq\alpha$ whenever $(a,u,v)\in C$. Lemmas 2.3 and 2.5 show that there exist a constant $M_1(a)$ such that, if $(a,u,v)\in C$, then $|u(x)|$, $|v(x)|<M_1(a)$ for all $x\in D$. Bootstrapping arguments imply that there exists a constant $M(a)$ such that $\|u\|$, $\|v\|<M(a)$ where $\|\ \|$ denotes the norm in $C^1(D)$. Hence the only way for C to approach ∞ in $\mathbb{R}\times C^1(D)\times C^1(D)$ is by a becoming unbounded. Since $\{a:(a,u,v)\in C\}$ is connected, it must equal $[\alpha,\infty)$.

Thus we obtain the following theorem on the existence of solutions of (1.5) to complement Theorem 5.2.

THEOREM 5.5. *Suppose* $b>\lambda_1-b_2/m$. *Then* (1.5) *has a nontrivial, nonnegative solution if and only if* $a>\alpha$ *where* $b=\lambda_1(-b_2q_\alpha)$ *i.e. if and only if* a *is sufficiently large so that* $b>\lambda_1(-b_2q_a)$.

REFERENCES

[1] H. AMANN, *Fixed point equations and nonlinear eigenvalue problems in ordered Banach spaces*, SIAM Rev., 18 (1976), pp. 620–709.

[2] J. BLAT AND K. J. BROWN, *Bifurcation of steady-state solutions in predator-prey and competition systems*, Proc. Roy. Soc. Edin. A, 97 (1984), pp. 21–34.

[3] R. S. CANTRELL AND C. COSNER, *On the steady-state problem for the Volterra–Lotka competition model with diffusion*, to appear.

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

[5] D. S. COHEN AND T. W. LAETSCH, *Nonlinear boundary value problems suggested by chemical reactor theory*, J. Differential Equations, 7 (1970), pp. 217–226.

[6] M. G. CRANDALL AND P. H. RABINOWITZ, *Bifurcation from simple eigenvalues*, J. Funct. Anal., 8 (1971), pp. 321–340.

[7] E. N. DANCER, *On positive solutions of some pairs of differential equations* I, Trans. Amer. Math. Soc., 284 (1984), pp. 729–743.

[8] M. A. KRASNOSELSKII, *Topological Methods in the Theory of Nonlinear Integral Equations*, Pergamon Press, Oxford, 1963.

[9] A. LEUNG, *Monotone schemes for semilinear elliptic systems related to ecology*, Math. Meth. Appl. Sci., 4 (1982), pp. 272–285.

[10] M. H. PROTTER AND H. F. WEINBERGER, *Maximum Principles in Differential Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1967.

[11] P. H. RABINOWITZ, *Some global results for non-linear eigenvalue problems*, J. Funct. Anal., 7 (1971), pp. 487–513.

[12] D. H. SATTINGER, *Topics in Stability and Bifurcation Theory*, Lecture Notes in Mathematics 309, Springer-Verlag, Heidelberg, 1973.

SIAM J. MATH. ANAL.
Vol. 17, No. 6, November 1986

005

SYMMETRY IN AN OVERDETERMINED FOURTH ORDER ELLIPTIC BOUNDARY VALUE PROBLEM*

ALLAN BENNETT†

Abstract. Let Ω be a bounded domain in $\mathbb{R}^N$ for which the following boundary value problem has a classical solution: $\Delta(\Delta u)=-1$ in Ω; $u=\partial u/\partial n$ on $\partial\Omega$; $\Delta u=c$ (constant) on $\partial\Omega$. We show that Ω must be an open ball and that u must be radially symmetric about the center of Ω. This result is analogous to that of Serrin (Arch. Rat. Mech. Anal., 43 (1971), pp. 304–318) and Weinberger (Arch. Rat. Mech. Anal., 43 (1971), pp. 319–320) for the problem $\Delta u=-1$ in Ω, $u=0$ and $\partial u/\partial n=c$ on $\partial\Omega$. Our result is obtained from a maximum principle for fourth order elliptic equations and several applications of Green's theorem. We then obtain two characterizations of open balls by means of integral identities—the first depends on our result and the second on that of Serrin and Weinberger.

Key words. maximum principle, radial symmetry

AMS(MOS) subject classification. Primary 35J

In 1971, J. Serrin [3] and H. Weinberger [4] proved that if Ω is a bounded domain in $\mathbb{R}^N$ with smooth boundary and if the solution to the problem

$$\Delta u=-1 \quad \text{in } \Omega, \tag{1a}$$

$$u=0 \quad \text{on } \partial\Omega \tag{1b}$$

has the property that $\partial u/\partial n$ is equal to a constant c on $\partial\Omega$, then Ω is a ball of radius $|Nc|$ and the solution to (1) is radially symetric about the center.

Serrin's proof is based on the Hopf maximum principle [2] and on a device of moving parallel planes to a critical position and then showing that the solution is symmetric about the limiting plane. This method can be extended to a more general second order elliptic problems. Weinberger's argument, however, is much more elementary. It also uses the maximum principle but relies on Green's theorem to establish certain identities which make it possible to solve for all second derivatives of the solution to (1). Unfortunately, Weinberger's argument does not extend to the more general results of [3]. However, the argument in [4] can be modified to establish the following theorem:

THEOREM. *Let Ω be a bounded domain in $\mathbb{R}^N$ with $C^{4+\varepsilon}$ boundary $\partial\Omega$, and suppose that the following overdetermined problem has a solution u in $C^4(\bar{\Omega})$:*

$$\Delta(\Delta u)=-1 \quad \textit{in } \Omega, \tag{2a}$$

$$u=\frac{\partial u}{\partial n}=0 \quad \textit{on } \partial\Omega, \tag{2b}$$

$$\Delta u\equiv c \quad \textit{on } \partial\Omega \ (c \textit{ constant}). \tag{2c}$$

Then Ω is an open ball of radius $[|c|(N^2+2N)]^{1/2}$, and

$$u(x)=\frac{-1}{2N}\left\{\frac{1}{4}(N+2)(Nc)^2+\frac{Nc}{2}r^2+\frac{1}{4(N+2)}r^4\right\}, \tag{2d}$$

where r denotes the distance from x to the center of Ω.

* Received by the editors January 22, 1985 and in final form September 30, 1985.
† Department of Mathematics, Cornell University, Ithaca, New York 14853.

This result has a corollary which allows a characterization of open balls in $\mathbb{R}^N$ by means of an integral identity:

COROLLARY. *Let D be a bounded domain in $\mathbb{R}^N$ with $C^{4+\varepsilon}$ boundary ∂D, and suppose that there is a real constant M so that*

$$\int_D B\,dx = M \oint_{\partial D} \frac{\partial B}{\partial n}\,dS \tag{3}$$

holds for any function B in $C^4(\bar{D})$ satisfying $\Delta(\Delta B)=0$ in D and $B=0$ on ∂D. Then D is an open ball, and

$$M = \frac{1}{|D|}\int_D |\nabla \psi|^2\,dx, \tag{4}$$

where $|D|$ is the N-dimensional volume of D and ψ satisfies $\Delta\psi = 1$ in D and $\psi = 0$ on ∂D.

Before proving these results, let us make some remarks about notation to be used. We shall use $u_{,i}$ and $u_{,ij}$ to denote, respectively, $\partial u/\partial x_i$ and $\partial^2 u/\partial x_i \partial x_j$. A superscript will denote a vector coordinate—for example, n^i is the ith coordinate of the (outward) unit normal vector to a surface at a given point. Finally, repeated indices indicate a summation with respect to that index, for example,

$$u_{,ii} = \Delta u = \sum_{i=1}^{N} \frac{\partial^2 u}{\partial x_i^2}.$$

Proof of Theorem. Suppose that $u \in C^4(\bar{\Omega})$ is a solution of the overdetermined problem (2). We first consider the following lemmas:

LEMMA 1 [1]. *The function*

$$\Phi := \frac{N-4}{N+2}u + \frac{N-4}{2(N+2)}(\Delta u)^2 + u_{,ij}u_{,ij} - \nabla u \cdot \nabla(\delta u) \tag{5}$$

assumes its maximum value on $\partial\Omega$.

LEMMA 2. *The following identity holds*:

$$\int_\Omega u\,dx = \frac{-N}{N+4}c^2|\Omega|. \tag{6}$$

To prove Lemma 1, it suffices to show that $\Delta\Phi \geqq 0$ in Ω. Routine calculation gives

$$\Delta\Phi = 2u_{,ijk}u_{,ijk} - \frac{6}{N+2}|\nabla(\Delta u)|^2. \tag{7}$$

To show that the right side of (7) is nonnegative, note that for any real number γ:

$$\sum_{i,j,k}\left[u_{,ijk} - \gamma\{(\Delta u)_{,i}\delta_{jk} + (\Delta u)_{,j}\delta_{ik} + (\Delta u)_{,k}\delta_{ij}\}\right]^2 \geqq 0 \tag{8}$$

or

$$u_{,ijk}u_{,ijk} - 6\gamma|\nabla(\Delta u)|^2 + 3\gamma^2(N+2)|\nabla(\Delta u)|^2 \geqq 0. \tag{9}$$

The discriminant of this quadratic expression in γ must then satisfy

$$36|\nabla(\Delta u)|^4 - 12(N+2)|\nabla(\Delta u)|^2 u_{,ijk}u_{,ijk} \leqq 0, \tag{10}$$

which is equivalent to $\Delta\Phi \geqq 0$. This yields the conclusion of Lemma 1.

The proof of Lemma 2 requires several applications of Green's theorem and the boundary condition in (2). First note that

$$\Delta\left[\Delta\left(r\frac{\partial u}{\partial r}\right)\right]=r\frac{\partial}{\partial r}[\Delta(\Delta u)]+4\Delta(\Delta u)=-4, \tag{11}$$

where r is the distance from a fixed origin. From (11) we obtain:

$$\begin{aligned}\int_\Omega\left[4u-r\frac{\partial u}{\partial r}\right]dx&=\int_\Omega\left[-u\Delta\Delta\left(r\frac{\partial u}{\partial r}\right)+r\frac{\partial u}{\partial r}\Delta\Delta u\right]dx \\ &=-\oint_{\partial\Omega}\Delta u\frac{\partial}{\partial n}\left(r\frac{\partial u}{\partial r}\right)dS=-c\oint_{\partial\Omega}\frac{\partial}{\partial n}\left(r\frac{\partial r}{\partial n}\frac{\partial u}{\partial n}\right)dS \\ &=-c\oint_{\partial\Omega}r\frac{\partial r}{\partial n}\frac{\partial^2 u}{\partial n^2}\,dS=-c^2\oint_{\partial\Omega}r\frac{\partial r}{\partial n}\,dS=-c^2N\cdot|\Omega|.\end{aligned} \tag{12}$$

An application of Green's theorem gives

$$\int_\Omega r\frac{\partial u}{\partial r}\,dx=-N\int_\Omega u\,dx, \tag{13}$$

so that (12) and (13) yield Lemma 2.

Our next step is to show that Φ is constant in Ω. We note that on $\partial\Omega$,

$$u_{,ij}u_{,ij}=\frac{\partial^2 u}{\partial n^2}n^in^j\frac{\partial^2 u}{\partial n^2}n^in^j=\left(\frac{\partial^2 u}{\partial n^2}\right)^2. \tag{14}$$

Lemma 1 and the boundary conditions (2) give

$$\Phi=\frac{3Nc^2}{2(N+2)}\quad\text{on }\partial\Omega,\qquad \Phi\leqq\frac{3Nc^2}{2(N+2)}\quad\text{in }\Omega. \tag{15}$$

By Lemma 2 and Green's theorem, we see that

$$\int_\Omega\Phi\,dx=\frac{-3(N+4)}{2(N+2)}\int_\Omega u\,dx=\frac{3Nc^2}{2(N+2)}\cdot|\Omega|. \tag{16}$$

Thus, by (15) and (16), $\Phi\equiv 3Nc^2/(2(N+2))$ in $\bar{\Omega}$. This implies that $\Delta\Phi$ vanishes identically in $\bar{\Omega}$. Therefore each term of the sum in (8) vanishes when $\gamma=1/(N+2)$. By differentiating each term with respect to x_k and adding, we obtain

$$u_{,ijkk}=(\Delta u)_{,ij}=\frac{1}{N+2}\left[2(\Delta u)_{,ij}-\delta_{ij}\right] \tag{17}$$

or

$$(\Delta u)_{,ij}=\frac{-1}{N}\delta_{ij}. \tag{18}$$

Now by (17) and analyticity of solutions of (2a), we have $\Delta u(x)=(1/2N)(A-|x-a|^2)$ in $\bar{\Omega}$, where A is the maximum value of Δu in $\bar{\Omega}$ and $\Delta u(a)=A$. By the boundary conditions (2c), we see that Ω is an open ball of radius $(A-2cN)^{1/2}$ centered at a. The boundary conditions (2b) then yield the radially symmetric solution (2d) and the additional relation $A=-cN^2$. This completes the proof of the theorem.

Proof of the corollary. Let $u \in C^4(\bar{D})$ be the solution of the problem

$$\Delta(\Delta u) = 1 \quad \text{in } D, \qquad u = \frac{\partial u}{\partial n} = 0 \quad \text{on } \partial D. \tag{19}$$

Then for any function B satisfying the hypothesis of the corollary, we have

$$M \int_{\partial D} \frac{\partial B}{\partial n}\, dS = \int_D B\, dx = \int_D \Delta(\Delta u) = -\oint_{\partial D} \Delta u \frac{\partial B}{\partial n}\, dS. \tag{20}$$

The last equality in (20) follows from several applications of Green's theorem and from the boundary value problems satisfied by u and B. We see now from (20) that

$$0 = \oint_{\partial D} \frac{\partial B}{\partial n}(M + \Delta u). \tag{21}$$

Let B be the solution of class $C^4(\bar{D})$ to the problem

$$\Delta(\Delta B) = 0 \quad \text{in } D, \qquad B = 0 \quad \text{and} \quad \frac{\partial B}{\partial n} = M + \Delta u \quad \text{on } \partial D. \tag{22}$$

It is immediate from (21) that $\Delta u = -M$ on ∂D, so that the theorem implies that D is an open ball. The exact value of M can be determined by replacing B in (3) by ψ in (4) and using an argument similar to the one above. Q.E.D.

To conclude this note, let us consider the analogue of the preceding corollary for harmonic functions. If D is an open ball in $\mathbb{R}^N$ and $h \in C^2(\bar{D})$ is harmonic in D, then the mean value theorem for harmonic functions implies that the average value of h over D and the average value of h on ∂D are both equal to the value of h at the center of D. Using the Serrin–Weinberger theorem, we can prove the following proposition: If Ω is a bounded domain in $\mathbb{R}^N$ with $C^{2+\varepsilon}$-boundary $\partial\Omega$ and if

$$\frac{1}{|\Omega|} \int_\Omega h = \frac{1}{|\partial\Omega|} \oint_{\partial\Omega} h \tag{23}$$

for each function in $C^2(\bar{\Omega})$ satisfying $\Delta h = 0$ in Ω, then Ω is an open ball. (Here, $|\partial\Omega|$ is the surface area of $\partial\Omega$.)

Proof. Let $u \in C^2(\bar{\Omega})$ satisfy $\Delta u = 1$ in Ω and $u = 0$ on $\partial\Omega$. Then:

$$\int_\Omega h = \int_\Omega h \Delta u = \int_\Omega u \Delta h + \oint_{\partial\Omega} \left[h \frac{\partial u}{\partial n} - u \frac{\partial h}{\partial n} \right] = \oint_{\partial\Omega} h \frac{\partial u}{\partial n} \tag{24}$$

$$= \frac{|\Omega|}{|\partial\Omega|} \oint_{\partial\Omega} h.$$

Therefore, we see that

$$0 = \oint_{\partial\Omega} h \left[\frac{\partial u}{\partial n} - \frac{|\Omega|}{|\partial\Omega|} \right]. \tag{25}$$

Now choose $h \in C^2(\bar{\Omega})$ so that $\Delta h = 0$ in Ω and $h = \partial u/\partial n - |\Omega|/|\partial\Omega|$ on $\partial\Omega$. Then (26) implies that $\partial u/\partial n| = |\Omega|/|\partial\Omega|$ on $\partial\Omega$, and the Serrin–Weinberger result completes the proof of the proposition. Q.E.D.

REFERENCES

[1] L. E. PAYNE (1976), *Some remarks on maximum principles*, J. Analyse Math., 30, pp. 421–33.
[2] M. PROTTER AND H. WEINBERGER (1967), *Maximum Principles in Differential Equations*, Prentice-Hall, Englewood Cliffs, NJ.
[3] J. SERRIN (1971), *A symmetry problem in potential theory*, Arch. Rat. Mech. Anal., 43, pp. 304–18.
[4] H. WEINBERGER (1971), *Remark on the preceding paper of Serrin*, Arch. Rat. Mech. Anal., 43, pp. 319–20.

SIAM J. MATH. ANAL.
Vol. 17, No. 6, November 1986

006

ON A SINGULAR NONLINEAR ELLIPTIC PROBLEM*

SÔNIA M. GOMES†

Abstract. Elliptic boundary value problems of the form $Lu=k(x)u^{-\alpha}$ in Ω and $u=0$ on the boundary of Ω are studied where L is given in the divergence form, $\alpha>0$ and $k(x)$ is a nonnegative measurable real function. Existence in $C^1(\bar{\Omega})$ and uniqueness of a solution $u\geqq 0$ are established for the equivalent fixed point problem $u(x)=\int_\Omega G(x,s)k(s)[u(s)]^{-\alpha}ds$, where $G(x,s)$ is the Green's function for the Dirichlet problem associated to L in Ω. Various inequalities for $G(x,s)$ are proved and a study is made of the action of the integral operator defined by the kernel G on unbounded functions with a prescribed growth near the boundary.

Introduction. This paper concerns positive solutions of elliptic boundary value problems of the form:

$$\text{(0.1)}\qquad \begin{aligned} &Lu(x)=-\sum_{i,j=1}^{n}\frac{\partial}{\partial x_i}\left(a_{ij}(x)\frac{\partial}{\partial x_j}u(x)\right)=f(x,u(x)) \quad \text{for } x\in\Omega,\\ &u(x)=0 \quad \text{for } x\in\partial\Omega, \end{aligned}$$

where Ω is a bounded region in R^n, $n\geqq 3$ and $\partial\Omega$ is the boundary of Ω. We will consider nonlinearities of the form:

$$f(x,u)=k(x)u^{-\alpha}, \qquad \alpha>0$$

where $k(x)$ is a nonnegative measurable real function.

The main point of study here, besides the existence and uniqueness, is the behavior of solutions near $\partial\Omega$, where f is singular.

We say that $u(x)$ is a solution of (0.1) if u satisfies

$$\text{(0.2)}\qquad u(x)=\int_\Omega G(x,s)k(s)[u(s)]^{-\alpha}ds$$

where $G(x,s)$ is the Green's function for the Dirichlet problem for the equation $Lu=h$ in Ω.

In §1 we establish various inequalities for $G(x,s)$. Using these inequalities, we will study in §2 the action of the integral operator defined by the kernel G on functions $h(x)$ which tend to infinity as $x\to\partial\Omega$ in a way we will make precise later.

The fixed point problem (0.2) may not have solutions in $H_0^1(\Omega)$. However, under appropriate assumptions, we will prove in §3 the existence and uniqueness of a solution $u(x)$ of (0.2), continuously differentiable in $\bar{\Omega}$. For this purpose, we seek a solution written in the form $u(x)=\Phi(x)\Psi(x)$, where $\Phi\in C^1(\bar{\Omega})\cap C^2(\Omega)$ satisfies $L\Phi(x)=1$ for $x\in\Omega$ and $\Phi(x)=0$ for $x\in\partial\Omega$. The new unknown $\Psi(x)$ must satisfy

$$\text{(0.3)}\qquad \Psi(x)=\int_\Omega \frac{G(x,s)k(s)}{\Phi(x)[\Phi(s)]^{\alpha}}[\Psi(s)]^{-\alpha}ds.$$

*Received by the editors March 30, 1984, and in revised form February 25, 1985.

†Instituto de Pesquisas Espaciais (INPE), Departamento de Energia Espacial, (12.201) São José dos Campos, S.P., Brazil.

Consequently, instead of (0.2), we have a new fixed point problem, that is, $\Psi = T \circ F(\Psi)$, where $F(\Psi)=[\Psi(x)]^{-\alpha}$ and T is the operator defined by the kernel

$$N(x,s)=\frac{G(x,s)k(s)}{\Phi(x)[\Phi(s)]^{\alpha}}.$$

We determine $\Psi \in C(\bar{\Omega})$, bounded away from zero.

With this process we are "avoiding the nonlinear singularity" but, on the other hand, we have to deal with an integral operator with a "bad" kernel. However, we still have that $T: C(\bar{\Omega}) \to C(\bar{\Omega})$ is a completely continuous operator.

By using this idea, in [5] we have already proved analogous results for the particular case of the sphere and $L=-\Delta$.

Singular differential equations like (0.1) arise in the theory of heat conduction in electrically conducting materials as discussed in [3]. In that paper the authors treat the existence question for the equation

$$u_t - \Delta u = f(x,t,u), \quad x \in \Omega \subset \mathbb{R}^n, \quad t>0,$$

coupled with initial and boundary conditions for a class of functions f which are nonincreasing in u. Assuming that $f(x,t,r) \to f(x,r)$ as $t \to \infty$, they obtain classical solutions of the corresponding elliptic boundary value problem upon letting $t \to \infty$.

Later, the existence of classical solutions was obtained in [9] for elliptic operators L more general than the Laplacian and for a class of functions $f(x,r)$ with no monotonicity assumption. The uniqueness was also obtained under appropriate assumptions.

In [2] the authors considered generalized solutions by means of a nonlinear eigenvalue problem. They assume the continuity of the coefficients of L and f, that $f(x,r)$ is bounded from above for $r>1$ and that $f(x,r) \to \infty$ as $r \to 0^+$, uniformly for $x \in \bar{\Omega}$. For $f(x,r)=f(r)$ they obtained bounds for the rate at which $u(x) \to 0$ when $x \to \partial\Omega$ and for $\|\operatorname{grad} u(x)\|$.

It is also interesting to observe that for $\Omega=(0,1) \subset \mathbb{R}$, the problem (0.1) appears in the study of similarity solutions of one-dimensional initial value problems for diffusion equations of the type $u_t=(k(x)|u_x|^{N-1}u_x)_x$, $N>0$ (cf. [1]).

Singular equations like (0.3) have been treated in [8] where the existence of solutions of

$$u(x)=\int_0^1 K(x,s)[u(s)]^{-\alpha} ds$$

is studied for $\alpha=1$ and for positive semidefinite symmetric kernels $K(x,s)$ satisfying $\int_0^1 K(x,s)\,ds \geqq \delta > 0$. That work was generalized for $\alpha>0$ in [6] but, in both papers, $K(x,s)$ is supposed to be continuous for $0 \leqq x,\ s \leqq 1$.

1. Inequalities for the Green's function. Let $\Omega \subset \mathbb{R}^n$, $n \geqq 3$ be a bounded region of class C^2. We consider a linear differential operator L defined by

$$Lu = -\sum_{i=1}^{n} \frac{\partial}{\partial x_i}\left(a_{ij}\frac{\partial}{\partial x_j}u\right),$$

where $a_{ij} \in C^{2,\lambda}(\bar{\Omega})$ for some $0<\lambda<1$ and $a_{ij}=a_{ji}$.

The operator L is assumed to be uniformly elliptic in the sense that there exists a constant $\gamma>0$ such that

$$\gamma \sum_{i=1}^{n} \xi_i^2 \leqq \sum_{i,j=1}^{n} a_{ij}(x)\xi_i\xi_j \leqq \gamma^{-1} \sum_{i=1}^{n} \xi_i^2$$

for all $x \in \bar{\Omega}$ and $\xi \in R^n$.

In what follows C is a generic constant which may be different at different places. The function $d(x)$ will denote the distance from x to $\partial\Omega$.

Let $G(x,y)$ be the Green's function of the Dirichlet problem for the equation $Lu=h$ in Ω. Under the above hypothesis the function G can be shown to exist with the following properties (cf. [7]):

(a) $G(x,y)$ is continuous in the variables x and y, x and y in Ω, with $x\neq y$, together with its first and second derivatives with respect to x_i.

(b) $L_x G(x,y)=0$ in Ω for $x\neq y$.

(c) $G(x,y)=0$ for $x\in\partial\Omega$.

(d) $G(x,y)=G(y,x)$.

Furthermore,

(1.1) (e) $G=O(\|x-y\|)^{2-n}$ uniformly in $\bar{\Omega}$.

First we will prove the following further properties of $G(x,y)$.

THEOREM 1.1. *For $x,y\in\Omega, x\neq y$, the Green's function $G(x,y)$ verifies*:

$$G(x,y)\leqq Cd(x)\|x-y\|^{1-n}, \tag{1.2}$$

$$\|\operatorname{grad}_x G(x,y)\|\leqq C\|x-y\|^{1-n}, \tag{1.3}$$

$$\|\operatorname{grad}_x G(x,y)\|\leqq Cd(y)\|x-y\|^{-n}, \tag{1.4}$$

where the constant C depends only on Ω and L.

In [10, Thm. 2.3], the author proves these inequalities for $L=-\Delta$ in Lyapunov–Dini regions. We will use the same idea of his proof.

Proof of Theorem 1.1. First we observe that if D is the region defined by

$$D=\left\{x=(x',x_n);\ x'\in R^{n-1},\ \|x'\|<1 \text{ and } -1+\sqrt{1-\|x'\|^2}<x_n<2\right\},$$

then there is some $s_0<1$ such that regions D_s congruent to D shrunk by a factor $1/s$ can be placed at every point $x_0\in\partial\Omega$ in such a way that the bent part intersects $\partial\Omega$ at this point only and the symmetry axis is along the normal, for all $s<s_0$.

Let $y\in\Omega$ be fixed. If $d(x)\geqq s_0$ we have

$$d(x)\geqq s_0\geqq Cd\geqq C\|x-y\|$$

where $d=$ diameter of Ω. Hence, by (1.1),

$$G(x,y)\leqq C|x-y|^{2-n}\leqq Cd(x)\|x-y\|^{1-n}.$$

The same type of argument holds if $d(x)<s_0$ but $\|x-y\|<2d(x)$. Thus it is sufficient to consider the case $d(x)<s_0$ and $d(x)<1/2\|x-y\|$. Let $x_0\in\partial\Omega$ be such that $d(x)=\|x-x_0\|$. At x_0 we place a region D_s as above, where

$$s=\begin{cases}\|x-y\|/4 & \text{if } \|x-y\|<4s_0,\\ s_0 & \text{otherwise.}\end{cases}$$

We may assume that $x_0=(0,s/2)$ and that the direction of the normal to Ω at x_0 is the x_n-axis. In this way $B\cap\bar{\Omega}=\{x_0\}$ where B is the ball $B_{s/2}(0)$.

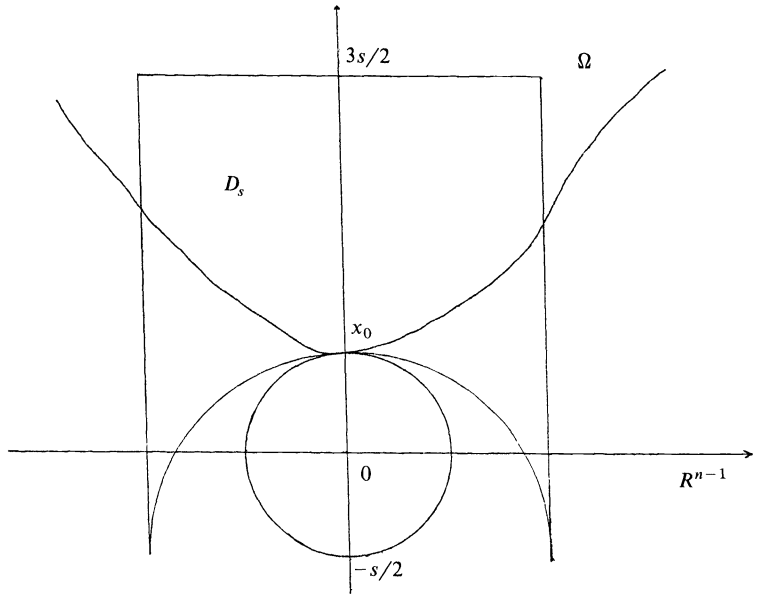


FIG. 1

We will construct "local barriers" at x_0 as follows. For suitable positive constants β_s and p, the function

$$w_s(x)=\beta_s\left[(s/2)^{-p}-\|x\|^{-p}\right]$$

satisfies:

$$\begin{aligned} &Lw_s\geqq 0 \quad \text{in } D_s\cap\Omega,\\ &w_s\geqq 1 \quad \text{in } \partial D_s\cap\Omega,\\ &w_s\geqq 0 \quad \text{in } \partial\Omega\cap D_s. \end{aligned} \tag{1.5}$$

Indeed:

$$Lw_s(x)=\beta_s p\|x\|^{-p-4}\left\{(p+2)\sum_{i,j=1}^{n}a_{ij}(x)x_ix_j - \|x\|^2\left[\sum_{i=1}^{n}a_{ii}(x)+\sum_{i,j=1}^{n}\frac{\partial}{\partial x_i}a_{ij}(x)x_j\right]\right\}.$$

Since the coefficients of L are bounded, there is a positive constant Λ such that

$$Lw_s(x)\geqq\beta_s p\|x\|^{-p-2}\left[(p+2)^{\gamma}-\Lambda(1+2s)\right].$$

Now, if p is sufficiently large so that $(p+2)\gamma-\Lambda(1+2s)\geqq 0$ for all $s\leqq s_0$, then $Lw_s(x)\geqq 0$ in $D_s\cap\Omega$ as asserted. If $x\in\partial D_s\cap\Omega$ then $\|x\|\geqq s$ and thus $w_s(x)\geqq \beta_s s^{-p}(2^p-1)=1$ provided $\beta_s=s^p/(2^p-1)$. If ν is the unit normal to Ω at the boundary point x_0 in the inward direction, then we also can see that

$$w_s(x_0+t\nu)\leqq Cts^{-1},$$

where the constant C is independent of s. Indeed,

$$w_s(x_0+t\nu)=\beta_s\left[(s/2)^{-p}-(t+s/2)^{-p}\right]=f_s(t);$$

$f_s'(t)=\beta_s p(t+s/2)^{-p-1}\geqq 0$ and $f_s''(t)=-\beta_s p(p+1)(t+s/2)^{-p-2}\leqq 0$. Hence $f_s(t)/t \leqq f_s'(0)=ps^{-p-1}2^{p+1}s^p/(2^p-1)=ps^{-1}2^{p+1}/(2^p-1)\leqq 4ps^{-1}$. It is easy to see that the distance from y to D_s is greater than $\|x-y\|/4$ which implies that

$$G(z,y)\leqq C\|x-y\|^{2-n} \quad \text{for } z\in\partial D_s\cap\Omega.$$

Now, we define $v(z)=w_s(z)-G(z,y)/C\|x-y\|^{2-n}$. In view of (1.5), we have

$$Lv=Lw_s\geqq 0 \quad \text{in } D_s\cap\Omega, \qquad v\geqq 0 \quad \text{in } \partial(D_s\cap\Omega).$$

Hence, $v(z)\geqq 0$ in $D_s\cap\Omega$. In particular, for $z=x=x_0+d(x)\nu$ we have that $w_s(x)\leqq Cd(x)s^{-1}$ and hence

$$G(x,y)\leqq C\|x-y\|^{2-n}w_s(x)\leqq C\|x-y\|^{2-n}d(x)s^{-1}.$$

If $\|x-y\|<4s_0$ then $s^{-1}=4/\|x-y\|$ and

$$G(x,y)\leqq Cd(x)\|x-y\|^{1-n}.$$

Otherwise,

$$G(x,y)\leqq Cd(x)\|x-y\|^{1-n}\|x-y\|/s_0\leqq Cd(x)\|x-y\|^{1-n},$$

and (1.2) is proved.

To prove (1.3), we consider first the points x such that $d(x)>\|x-y\|$. Let $B=B_r(x)$ where $r=\|x-y\|/4$. Since $G(x,y)$ belongs to $C^{2,\lambda}(B)$ (cf. [4, Thm. 6.13]) and in view of [4, Cor. 6.3], we can assert that

$$\|x-y\|\|\operatorname{grad}_x G(x,y)\|\leqq C\sup_{z\in B}G(z,y)$$

where C depends on the ellipticity constant γ and on the $C^{\lambda}(\bar{\Omega})$ bounds on the coefficients of L (as well as on n).

Hence,

$$\begin{aligned}\|\operatorname{grad}_x G(x,y)\|&\leqq C\|x-y\|^{-1}\sup_{z\in B}\|z-y\|^{2-n}\\ &\leqq C\|x-y\|^{-1}\|x-y\|^{2-n}\\ &=C\|x-y\|^{1-n}.\end{aligned}$$

If $d(x)\leqq\|x-y\|$ we consider $B=B_r(x)$ with $r=d(x)/4$. By the same argument we have

$$\begin{aligned}\|\operatorname{grad}_x G(x,y)\|&\leqq C[d(x)]^{-1}\sup_{z\in B}G(z,y)\\ &\leqq C[d(x)]^{-1}\sup_{z\in B}d(z)\|z-y\|^{1-n}\\ &\leqq C\sup_{z\in B}\|z-y\|^{1-n}\\ &\leqq C\|x-y\|^{1-n}.\end{aligned}$$

To prove (1.4), we first show that for fixed $x \in \Omega$

$$\frac{\partial}{\partial x_i} G(x,y) \to 0 \quad \text{as } y \to \partial\Omega.$$

Let $\rho > 0$, small enough such that

$$G(x,z) \geqq \frac{1}{2} \|x-z\|^{2-n} \quad \text{for } \|x-z\| = \rho.$$

Now, let $t \in \mathbb{R}$ with $|t| < \rho/2$. Hence, for $\|x-z\| = \rho$ and for some $\tilde{x}$ in the segment $[x, x+te_i]$ we have

$$\left| \frac{G(x+te_i,z) - G(x,z)}{t} \right| = \|\operatorname{grad}_x G(\tilde{x},z)\| \leqq C\|\tilde{x}-z\|^{1-n}$$

$$\leqq C\|x-z\|^{1-n}$$

$$\leqq C\rho^{-1} G(x,z).$$

Since this difference quotient is zero for $z \in \partial\Omega$, this inequality holds also for $z \in \Omega - B$ where $B = B_\rho(x)$. Therefore, if $y \in \Omega - B$ we have that

$$|\partial/\partial x_i G(x,y)| = \lim_{t \to 0} \left| \frac{G(x+te_i,y) - G(x,y)}{t} \right|$$

$$\leqq C\rho^{-1} G(x,y).$$

Hence

$$\frac{\partial}{\partial x_i} G(x,y) \to 0 \quad \text{as } d(y) \to 0.$$

Next, since $L_y \partial/\partial x_i G(x,y) = 0$ and using (1.3), the inequality (1.4) follows as in the proof of (1.2).

2. Consider the integral operator defined by the kernel $G(x,y)$. The proof of the following lemma is analogous to that of [4, Lemma 4.1] and it is a consequence of (1.1) and (1.3).

LEMMA 2.1. *Let $h : \Omega \to R$ be a bounded measurable function. If $v(x)$ is the function defined by*

$$v(x) = \int_\Omega G(x,s) h(s)\, ds,$$

then

$$v \in C^1(\bar{\Omega}),$$

$$\frac{\partial}{\partial x_i} v(x) = \int_\Omega \frac{\partial}{\partial x_i} G(x,s) h(s)\, ds \quad \textit{for all } x \in \bar{\Omega}.$$

As we said in the introduction, we will study here the behavior of the function $v(x)$ when the function h goes to infinity on the boundary. More precisely, the inequalities (1.2)–(1.4) will allow us to consider unbounded functions h with the following condition on their growth near $\partial\Omega$.

(2.1) There is some $0 \leqq \tau < 1$ such that $[d(x)]^\tau h(x) \in L^\infty(\Omega)$.

THEOREM 2.2. *If* $h:\Omega\to R$ *is a measurable function that satisfies* (2.1) *and* $v(x)=\int_\Omega G(x,s)h(s)\,ds$, *then*:

$$v\in C^1(\bar\Omega), \tag{2.2}$$

$$\frac{\partial}{\partial x_i}v(x)=\int_\Omega \frac{\partial}{\partial x_i}G(x,s)h(s)\,ds \quad \textit{for all } x\in\bar\Omega. \tag{2.3}$$

Proof of Theorem 2.2. By virtue of (1.1) the function v is well defined. In what follows, we will use several times the inequalities (1.2), (1.3) and (1.4). Let $w(x)=\int_\Omega \partial/\partial x_i G(x,s)h(s)\,ds$. First we will prove that w is well defined for $x\in\bar\Omega$.

(i) If $x\in\Omega$ and $r<d(x)$, we have:

$$|w(x)|\leqq C\left[\int_{\|x-s\|<r}\|x-s\|^{1-n}|h(s)|\,ds+r^{1-n}\int_{\|x-s\|\geqq r}|h(s)|\,ds\right]$$

$$\leqq C\left[\int_{\|x-s\|<r}\|x-s\|^{1-n}ds+r^{1-n}\int_{\|x-s\|\geqq r}[d(s)]^{-\tau}ds\right]<\infty.$$

(ii) If $x\in\partial\Omega$ and $r>0$, we have:

$$|w(x)|\leqq C\left[\int_{\|x-s\|<r}[d(s)]^{1-\tau}\|x-s\|^{-n}ds+\int_{\|x-s\|\geqq r}[d(s)]^{-\tau}\|x-s\|^{1-n}ds\right]$$

$$\leqq C\left[\int_{\|x-s\|<r}\|x-s\|^{1-n-\tau}ds+r^{1-n}\int_{\|x-s\|\geqq r}[d(s)]^{-\tau}ds\right]<\infty.$$

Next, we define $h_\varepsilon(x)=h(x)\eta(d(x)/\varepsilon)$ where η is the characteristic function of the interval $[1,+\infty)$; that is, $\eta(t)=1$ if $t\geqq 1$ and $\eta(t)=0$ if $t<1$. Now h_ε is bounded and thus, the function

$$v_\varepsilon(x)=\int_\Omega G(x,s)h_\varepsilon(s)\,ds$$

is in $C^1(\bar\Omega)$ and $\partial/\partial x_i v_\varepsilon(x)=\int_\Omega \partial/\partial x_i G(x,s)h_\varepsilon(s)\,ds=w_\varepsilon(x)$ for all $x\in\bar\Omega$.

$$v(x)-v_\varepsilon(x)=\int_{\Omega_\varepsilon}G(x,s)h(s)\,ds \quad \text{where } \Omega_\varepsilon=\{x\in\Omega;\ d(x)<\varepsilon\}.$$

Hence,

$$|v(x)-v_\varepsilon(x)|\leqq\int_{\Omega_\varepsilon}G(x,s)|h(s)|\,ds\leqq\int_{\Omega_\varepsilon}[d(s)]^{1-\tau}\|x-s\|^{1-n}ds$$

$$\leqq C\varepsilon^{1-\tau}.$$

Now,

$$w(x)-w_\varepsilon(x)=\int_{\Omega_\varepsilon}\frac{\partial}{\partial x_i}G(x,s)h(s)\,ds.$$

Thus,

(i) If $d(x)>2\varepsilon$ and $s\in\Omega_\varepsilon$ then $\|x-s\|\geqq\varepsilon>d(s)$ and

$$|w(x)-w_\varepsilon(x)|\leqq C\int_{\Omega_\varepsilon}[d(s)]^{1-\tau}\|x-s\|^{-n}ds$$

$$\leqq C\int_{\Omega_\varepsilon}\|x-s\|^{1-n-\tau}ds$$

$$\leqq C\left[\int_{\|x-s\|<d}\|x-s\|^{p(1-n-\tau)}ds\right]^{1/p}\mu(\Omega_\varepsilon)^{1/q}$$

where $p>1$ is such that $p(n-1+\tau)<n$, $q=p/(p-1)$, $d=$ diameter of Ω and μ is the Lebesgue measure.

(ii) If $d(x)\leqq\varepsilon$ we shall consider a partition of Ω_ε as follows:

$$A_1=\{s\in\Omega_\varepsilon;\ \|x-s\|<2d(x) \text{ and } d(s)<\tfrac{1}{2}d(x)\},$$
$$A_2=\{s\in\Omega_\varepsilon;\ \|x-s\|<2d(x) \text{ and } d(s)\geqq\tfrac{1}{2}d(x)\}, \quad \text{and}$$
$$A_3=\Omega_\varepsilon-(A_1\cup A_2).$$

In this way we have:

$$|w(x)-w_\varepsilon(x)|\leqq I_1+I_2+I_3 \quad \text{where } I_j=\int_{A_j}|\partial/\partial x_i G(x,s)h(s)|\,ds.$$

$$I_1\leqq C[d(x)]^{1-n-\tau}\int_{A_1}ds\leqq C[d(x)]^{1-\tau}\leqq C\varepsilon^{1-\tau},$$

$$I_2\leqq C\int_{A_2}[d(s)]^{-\tau}\|x-s\|^{1-n}ds\leqq C[d(x)]^{-\tau}\int_{A_2}\|x-s\|^{1-n}ds\leqq C[d(x)]^{1-\tau}$$
$$\leqq C\varepsilon^{1-\tau},$$

$$I_3\leqq C\int_{A_3}[d(s)]^{1-\tau}\|x-s\|^{-n}ds\leqq C\int_{A_3}\|x-s\|^{1-n-\tau}ds$$
$$\leqq C\left[\int_{\|x-s\|<d}\|x-s\|^{p(1-n-\tau)}ds\right]^{1/p}\mu(\Omega_\varepsilon)^{1/q}.$$

For the last inequality we used the fact that if $s\in A_3$, that is $\|x-s\|\geqq 2d(x)$, then

$$d(s)=\|s-s_0\|\leqq\|s-x_0\|\leqq\|x-s\|+\|x-x_0\|=\|x-s\|+d(x)$$
$$\leqq\|x-s\|+1/2\|x-s\|=3/2\|x-s\|,$$

where x_0 and s_0 are in $\partial\Omega$.

So, we have proved that $|v(x)-v_\varepsilon(x)|\leqq C\varepsilon^{1-\tau}$ and $|w(x)-w_\varepsilon(x)|\leqq C\mu(\Omega_\varepsilon)^{1/q}$. Consequently v_ε and w_ε converge to v and to w respectively as $\varepsilon\to 0$. Hence, $v\in C^1(\bar{\Omega})$ and $\partial/\partial x_i v=w$ and the theorem is proved.

Remark. The first assertion of Theorem 2.2 above is well known under more restrictive assumptions; for example if $h\in L^p(\Omega)$, $p>n$. Assuming also that h satisfies the condition (2.1) the Holder continuity of the grad v is stated in [10] for $v\in C^2(\Omega)$. However, in the proof of Theorem 3.1 below, we will need, not only the regularity condition (2.2), but also an integral representation for grad v, as stated in (2.3).

3. The main theorem. In view of Theorem 2.2 above, we observe that if $u(x)$ is a solution of (0.2) and $k(x)[u(x)]^{-\alpha}$ verifies the condition (2.1) then $u \in C^1(\bar{\Omega})$. In this case, since $\partial/\partial\nu u(x) > 0$ for all $x \in \partial\Omega$, there are positive numbers θ and Θ such that $\theta d(x) \leqq u(x) \leqq \Theta d(x)$. Consequently, $k(x)[d(x)]^{-\alpha} \leqq Ck(x)[u(x)]^{-\alpha}$ and $k(x)[d(x)]^{-\alpha}$ also verifies (2.1). Conversely, we will prove in the present section the existence and uniqueness of a solution $u(x)$ of (0.2), $u(x)$ continuously differentiable in $\bar{\Omega}$, provided that $k(x)[d(x)]^{-\alpha}$ verifies (2.1).

Let $\Phi \in C^1(\bar{\Omega}) \cap C^2(\Omega)$ be the solution of

$$\begin{aligned} L\Phi(x) &= 1 \quad \text{for } x \in \Omega, \\ \Phi(x) &= 0 \quad \text{for } x \in \partial\Omega. \end{aligned}$$

We observe that $\Phi(x) = \int_\Omega G(x,s)\,ds$ and that $\partial/\partial\nu\Phi(x) > 0$ for all $x \in \partial\Omega$.

As mentioned in the introduction, we will seek a solution in the form $u(x) = \Phi(x)\Psi(x)$ where $\Psi(x)$ must satisfy (0.3); that is, the fixed point problem $\Psi = T \circ F(\Psi)$ where $F(\Psi)(x) = [\Psi(x)]^{-\alpha}$ and

$$Tw(x) = \int_\Omega \frac{G(x,s)k(s)}{\Phi(x)[\Phi(s)]^\alpha} w(s)\,ds.$$

We will determine the existence and uniqueness of the fixed point $\Psi \in \mathring{K}$ where $\mathring{K} = \{v \in C(\bar{\Omega});\ v(x) > 0 \text{ for all } x \in \bar{\Omega}\}$ is the cone of the positive functions. For this purpose we will prove first the following theorem:

THEOREM 3.1. *If $k(x)[d(x)]^{-\alpha}$ verifies the condition* (2.1), *then $T: C(\bar{\Omega}) \to C(\bar{\Omega})$ is a completely continuously linear operator that preserves $\mathring{K}$.*

Proof of Theorem 3.1. The complete continuity of the operator T in $C(\bar{\Omega})$ is satisfied if it can be represented in the form

$$Tw(x) = \int_\Omega N(x,s)w(s)\,ds$$

where the kernel function $N(x,s)$ verifies the following condition:

$$\lim_{y \to x} \int_\Omega |N(x,s) - N(y,s)|\,ds = 0 \quad \text{for all } x \in \bar{\Omega}. \tag{3.1}$$

For this purpose we define for $s \in \Omega$

$$N(x,s) = \begin{cases} \dfrac{G(x,s)k(s)}{\Phi(x)[\Phi(s)]^\alpha} & \text{for } x \in \Omega, \\ \dfrac{\partial/\partial\nu G(x,s)k(s)}{\partial/\partial\nu\Phi(x)[\Phi(s)]^\alpha} & \text{for } x \in \partial\Omega. \end{cases}$$

LEMMA 3.2. *Under the hypothesis of Theorem* 3.1 *and for all $x \in \bar{\Omega}$ we have*:

$$\lim_{y \to x} N(y,s) = N(x,s), \tag{3.2}$$

$$\lim_{y \to x} \int_\Omega N(y,s)\,ds = \int_\Omega N(x,s)\,ds. \tag{3.3}$$

Proof of Lemma 3.2. Since $\Phi \in C^2(\Omega)$ and $G \in C^2(\Omega - \{s\})$, we need to verify (3.2) only for $x \in \partial\Omega$. By virtue of the continuity of grad Φ and grad G up to the boundary of Ω, we have:

$$G(x+t\nu,s) = \frac{\partial}{\partial \nu} G(x,s)t + E_1(x,t);$$

$$\Phi(x+t\nu) = \frac{\partial}{\partial \nu}\Phi(x)t + E_2(x,t),$$

where $\lim_{t\to 0} E_i(x,t)/t = 0$ uniformly for $x \in \partial\Omega$. Therefore, given $\varepsilon > 0$ there exists $\delta_1 > 0$ such that for all $x \in \partial\Omega$ and for $0 \leqq t \leqq \delta_1$:

$$|N(x+t\nu,s) - N(x,s)|$$

$$= k(s)|\Phi(s)|^{-\alpha} \left| \frac{\partial/\partial\nu\Phi(x) E_1(x,t)/t - \partial/\partial\nu G(x,s) E_2(x,t)/t}{|\partial/\partial\nu\Phi(x)|^2 + \partial/\partial\nu\Phi(x) E_2(x,t)/t} \right|$$

$$< \varepsilon/2.$$

Since the direction of the normals is a continuous function on $\partial\Omega$, there exists $\delta_2 > 0$ such that $|N(y,s) - N(x,s)| < \varepsilon/2$ for x and $y \in \partial\Omega$ with $\|x-y\| < \delta_2$. Hence, given $x \in \partial\Omega$ and $\varepsilon > 0$, let $y \in \bar{\Omega}$ be such that $\|x-y\| < \delta$, where $\delta = \min\{\delta_1, \delta_2/4\}$. If $\bar{y} \in \partial\Omega$ verifies $d(y) = \|y - \bar{y}\|$ then,

$$\|x - \bar{y}\| \leqq \|x-y\| + \|y-\bar{y}\| \leqq 2\|x-y\| < 2\delta \leqq \delta_2/2 < \delta_2.$$

Thus,

$$|N(x,s) - N(y,s)| \leqq |N(x,s) - N(\bar{y},s)| + |N(\bar{y},s) - N(y,s)| < \varepsilon.$$

To prove (3.3), we proceed as in the proof of (3.2) observing that $\int_\Omega N(x,s)\,ds = \Gamma(x)/\Phi(x)$ for $x \in \Omega$ and $\int_\Omega N(x,s)\,ds = \partial/\partial\nu\Gamma(x)/\partial/\partial\nu\Phi(x)$ for $x \in \partial\Omega$, where

$$\Gamma(x) = \int_\Omega \frac{G(x,s)k(s)}{[\Phi(s)]^\alpha} ds \in C^1(\bar{\Omega}).$$

Completion of the proof of Theorem 3.1. Now, (3.1) follows as a consequence of (3.2) and (3.3) and the fact that if the functions g and $g_m \in L^1(\Omega)$ are such that $g_m(x) \to g(x)$ a.e. and $\int_\Omega |g_m(s)|\,ds \to \int_\Omega |g(s)|\,ds$ then $\int_\Omega |g_m(s) - g(s)|\,ds \to 0$. To conclude the proof of Theorem 3.1, we consider $w(x) \in \mathring{K}$. $Tw(x) = \int_\Omega N(x,s)w(s)\,ds$ is a continuous function in $\bar{\Omega}$, $Tw(x) \geqq 0$ and $Tw(x) = 0$ if and only if $N(x,s) \equiv 0$. But this takes place only if $k(x) \equiv 0$. So, Tw is a strictly positive continuous function in $\bar{\Omega}$, that is, $Tw \in \mathring{K}$, and the Theorem 3.1 is proved.

Now, for $0 < \varepsilon \leqq 1$ we define $f_\varepsilon(t) = \varepsilon^{-\alpha}$ for $t \leqq \varepsilon$ and $f_\varepsilon(t) = t^{-\alpha}$ for $t > \varepsilon$. If $F_\varepsilon : C(\bar{\Omega}) \to C(\bar{\Omega})$ is such that $F_\varepsilon w(x) = f_\varepsilon(w(x))$, then it follows immediately from the Theorem 3.1 above, that $T \circ F_\varepsilon : \mathring{K} \to \mathring{K}$ is a continuous and compact nonlinear operator. Under these conditions we will prove:

LEMMA 3.3. *There exists a unique function* $\Psi_\varepsilon \in \mathring{K}$ *such that*

$$T \circ F_\varepsilon(\Psi_\varepsilon) = \Psi_\varepsilon.$$

Furthermore, if $0 < \varepsilon \leqq \bar{\varepsilon} \leqq 1$ *then* $\Psi_{\bar{\varepsilon}} \leqq \Psi_\varepsilon$.

Proof of Lemma 3.3. Let $R \geqq 1$ be a number such that $1/R \leqq \int_\Omega N(x,s)\,ds \leqq R$, for all $x \in \bar{\Omega}$. For all $w \in \mathring{K}$, $T \circ F_\varepsilon(w) \leqq R\varepsilon^{-\alpha}$. If we define $w_2(x) = R\varepsilon^{-\alpha}$, then $T \circ F_\varepsilon(w_2) \geqq R^{-1-\alpha}\varepsilon^{\alpha^2}$, and if $w_1(x) = R^{-1-\alpha}\varepsilon^{\alpha^2}$, then $T \circ F_\varepsilon(w_1) \leqq w_2$. So, $T \circ F_\varepsilon : [w_1, w_2] \to [w_1, w_2]$ where $[w_1, w_2] = \{w \in C(\bar{\Omega});\ w_1(x) \leqq w(x) \leqq w_2(x) \text{ for all } x \in \bar{\Omega}\}$. Now, the existence of

$\Psi_\varepsilon \in [w_1, w_2]$, fixed point of $T \circ F_\varepsilon$, is a consequence of Schauder's Theorem. The uniqueness of Ψ_ε is a consequence of the Maximum Principle (cf. [4, Thm. 8.8]). Indeed, if $\Psi_{\varepsilon,1}$ and $\Psi_{\varepsilon,2}$ are two fixed points and $h(x) = \Phi(x)\Psi_{\varepsilon,1}(x) - \Phi(x)\Psi_{\varepsilon,2}(x)$, then $h \in H_0^1(\Omega)$ (in fact, $h \in C^1(\bar{\Omega})$, cf. Theorem 2.2) and

$$Lh(x) = k(x)[\Phi(x)]^{-\alpha}[f_\varepsilon(\Psi_{\varepsilon,1}(x)) - f_\varepsilon(\Psi_{\varepsilon,2}(x))].$$

Let A be the open set $\{x \in \Omega;\ \Psi_{\varepsilon,1}(x) > \Psi_{\varepsilon,2}(x)\}$. For $x \in \partial A$ $h(x) = 0$ and $Lh(x) \leqq 0$ in A. Hence, $h(x) \leqq 0$ in A. But this is a contradiction, unless $A = \varnothing$.

In the same way, if $0 < \varepsilon \leqq \bar{\varepsilon} \leqq 1$, $A = \{x \in \Omega;\ \Psi_\varepsilon(x) > \Psi_{\bar{\varepsilon}}(x)\}$ and $h(x) = \Phi(x)\Psi_{\bar{\varepsilon}}(x) - \Phi(x)\Psi_\varepsilon(x)$ then $h \in H_0^1(\Omega)$ and

$$\begin{aligned} Lh(x) &= k(x)[\Phi(x)]^{-\alpha}[f_{\bar{\varepsilon}}(\Psi_{\bar{\varepsilon}}(x)) - f_\varepsilon(\Psi_\varepsilon(x))] \\ &\leqq k(x)[\Phi(x)]^{-\alpha}[f_\varepsilon(\Psi_{\bar{\varepsilon}}(x)) - f_\varepsilon(\Psi_\varepsilon(x))]. \end{aligned}$$

Hence, $h(x) = 0$ for $x \in \partial A$ and $Lh(x) \leqq 0$ in A and so, $h(x) \leqq 0$ in A which is a contradiction, unless $A = \varnothing$.

We are now in a position to prove our main theorem:

THEOREM 3.4. *If $k(x)[d(x)]^{-\alpha}$ verifies the condition* (2.1), *then the problem* (0.2) *has a unique solution u and $u \in C^1(\bar{\Omega})$.*

Proof of Theorem 3.4. As a consequence of Lemma 3.3 above we have that $\Psi_\varepsilon(x) \geqq \Psi_1(x) \geqq R^{-1-\alpha}$ for $0 < \varepsilon \leqq 1$. Therefore, if $\varepsilon = R^{-1-\alpha}$ and $\Psi = \Psi_\varepsilon$, then $F_\varepsilon(\Psi) = F(\Psi)$ and $T \circ F(\Psi) = T \circ F_\varepsilon(\Psi) = \Psi$. Now, if $u(x) = \Phi(x)\Psi(x)$ then u verifies (0.2). The uniqueness of u follows from the uniqueness of Ψ. In view of Theorem 2.2 we conclude that $u \in C^1(\bar{\Omega})$.

Acknowledgment. The author wishes to thank Professor Pedro Nowosad from Instituto de Matemática Pura e Aplicada(CNPq)–Rio de Janeiro, for suggesting the present problem and for the encouragement given to her.

REFERENCES

[1] J. E. BOUILLET AND S. M. GOMES, *An equation with a singular nonlinearity related to diffusion problems in one dimension*, Quart. Appl. Math., to appear.

[2] M. G. CRANDALL, P. H. RABINOWITZ AND L. TARTAR, *On a Dirichlet problem with a singular nonlinearity*, Comm. Partial Differentiation Equations, 2 (1977),

[3] W. FULKS AND J. S. MAYBEE, *A singular nonlinear equation*, Osaka J. Math., 12 (1960), pp. 1–19.

[4] D. GILBARG AND N. S. TRUDINGER, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, New York, 1977.

[5] S. M. GOMES, *Existência e comportamento na fronteira de solução de problema de Dirichlet não linear singular na fronteira*, Ph.D. dissertation, Instituto de Matemática Pura e Aplicada–CNPq, Rio de Janeiro, Brasil, 1982.

[6] S. KARLIN AND L. NIRENBERG, *On a theorem of P. Nowosad*, J. Math. Anal. Appl., 17 (1967), pp. 61–67.

[7] C. MIRANDA, *Partial Differential Equations of Elliptic Type*, Springer-Verlag, New York, 1970.

[8] P. NOWOSAD, *On the integral equation $kf = 1/f$ arising in a problem in communications*, J. Math. Anal. Appl., 14 (1966), pp. 484–492.

[9] C. A. STUART, *Existence and approximation of solutions of nonlinear elliptic equations*, Math. Z., 147 (1976), pp. 53–63.

[10] K. O. WIDMAN, *Inequalities for the Green function and boundary continuity of the gradient of solutions of elliptic differential equations*, Math. Scand., 21 (1967), pp. 17–37.

SIAM J. MATH. ANAL.
Vol. 17, No. 6, November 1986

007

ELLIPTIC PROBLEMS ON UNBOUNDED DOMAINS*

RAINER JANßEN†

Abstract. This paper considers elliptic boundary value problems on unbounded domains with possibly unbounded boundary using the variational method. Since a nonvanishing harmonic function is not square integrable on R^n, the construction of solutions in the usual Sobolev spaces, so successful for bounded domains, must fail. Kudrjavcev [8], [9] showed how to circumvent this difficulty by introducing weighted Sobolev spaces. See Besov et al. [3] for a survey on this method. Benci, Fortunato [2], Cantor [4], Janßen [5], Mäulen [10], Owen [15], Vogelsang [16], and others, apparently independently, have seized upon this method for treating elliptic problems on unbounded domains. This paper generalizes the method in two directions: 1) We treat operators of all orders with Dirichlet, Neumann and mixed boundary conditions. 2) We impose none of the usual restrictions on the coefficients of the operator (e.g. that they should converge to a constant as $|x|\to\infty$). The main tools are variants of the Poincaré and Friedrichs inequality, respectively, and compact embeddings in weighted Sobolev spaces.

Key words. elliptic boundary value problems, unbounded domains, weighted Sobolev spaces

AMS(MOS) subject classifications. Primary 35J20, 26A86

1. Introduction and notation.

As mentioned in the abstract, we consider elliptic boundary value problems in weighted Sobolev spaces. Before we give the definition of these spaces let us fix some notation.

From now on let Ω denote an open set in R^n with boundary $\partial\Omega$. Furthermore, let

$$C_0^\infty := \{f:\Omega\to R \mid f \text{ infinitely often differentiable with compact support in } \Omega\},$$

$$C^\infty := \{f:\Omega\to R \mid f \text{ infinitely often differentiable in } \Omega\},$$

$$D_i f := \frac{\partial f}{\partial x_i}, \qquad i=1,\cdots,n,$$

$$D^\alpha f := \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1}\cdots\partial x_n^{\alpha_n}}, \quad \text{where } \alpha=(\alpha_1,\cdots,\alpha_n)\in(N_0)^n, \quad |\alpha|=\sum_{i=1}^n \alpha_i.$$

A positive-real valued, function ρ is called a weight function. If ρ is a weight function and $k\in N$, then

$$(f,g)_{k,\rho} := \sum_{|\alpha|\leqq k}\int_\Omega \rho^{-2k+2|\alpha|} D^\alpha f D^\alpha g\, dx$$

defines a scalar product on $C_0^\infty(\Omega)$. The completion with respect to the appropriate norm is called $\overset{0}{W}{}^{k,\rho}(\Omega)$; hence

$$\overset{0}{W}{}^{k,\rho}(\Omega) := \overline{C_0^\infty(\Omega)}^{\|\cdot\|_{k,\rho}}.$$

Similarly we define

$$W_c^{k,\rho}(\Omega) := \overline{C_0^\infty(R^n)|_\Omega}^{\|\cdot\|_{k,\rho}},$$

$$W^{k,\rho}(\Omega) := \overline{C^\infty}^{\|\cdot\|_{k,\rho}},$$

*Received by the editors December 19, 1984, and in revised form June 30, 1985.
† Wissenschaftliches Zentrum der IBM, Tiergartenstrasse 15, 6900 Heidelberg, West Germany.

where $C_0^\infty(R^n)|_\Omega$ denotes the space of restrictions of functions in $C_0^\infty(R^n)$ to Ω. For $x=(x_1,\cdots,x_n)\in R^n$ we need the following three norms

$$|x|_1=\sum_1^n |x_i|, \quad |x|_2=\left(\sum_1^n x_i^2\right)^{1/2}, \quad |x|_\infty=\max_{i=1,\cdots,n}|x_i|.$$

$L_q(\Omega)$ denotes the usual Lebesgue space with $\|f\|_q := (\int_\Omega |f|^q dx)^{1/q}$.

Remark 1.1. It is clear that $\overset{0}{W}{}^{k,\rho}\subset W_c^{k,\rho}\subset W^{k,\rho}$. The functions in $W_c^{k,\rho}$ need not vanish on the boundary as the functions in $\overset{0}{W}{}^{k,\rho}$ but contrary to $W^{k,\rho}$ there is a kind of growth condition at infinity.

Let us now sketch the plan of the paper. In the next section we prove variants of the Poincaré lemma and apply this to study the Dirichlet problem. The third section is devoted to compact imbeddings and their applications. In the last section we shall study the Neumann problem. There we shall again prove suitable variants of the Poincaré lemma, this time for functions which do not necessarily vanish on the boundary. In most existence theorems we shall get explicit bounds for the norm of the solution operator.

2. Dirichlet problem on unbounded domains. The Poincaré lemma states that for bounded domains Ω there exists a constant depending only on the diameter of Ω such that for all $f\in C_0^\infty(\Omega)$ $\int f^2 dx \leqq \text{const}\int |\nabla f|^2 dx$. It is easy to see that this is false, in general, for unbounded domains. We prove now some variants for the weighted spaces.

LEMMA 2.1. *Let ρ be a locally Lipschitz-continuous weight function with $\rho(x)\geqq|x|_\infty$ on Ω. Furthermore, let $k\in N$ and $q,\delta\in R$ with $q>1$ and $\delta>0$. Then either of the two conditions*

$$\text{(i)} \qquad 1-\frac{qk\sum_1^n D_i\rho(x)x_i}{n\rho(x)}\geqq\delta \quad \textit{for all } x\in\Omega,$$

$$\text{(ii)} \qquad 1-\frac{qk\sum_1^n D_i\rho(x)x_i}{n\rho(x)}\leqq-\delta \quad \textit{for all } x\in\Omega$$

implies that for all $f\in C_0^\infty(\Omega)$

$$\|f\rho^{-k}\|_q\leqq\frac{q}{n\delta}\sum_1^n\|(D_if)\rho^{-k+1}\|_q.$$

Proof. Integration by parts implies

$$\int\left|\frac{f}{\rho^k}\right|^q dx=+\frac{1}{n}\sum_1^n\int\left|\frac{f}{\rho^k}\right|^q dx=-\frac{1}{n}\sum_1^n\int\operatorname{sign}(f)D_i\left(\frac{f}{\rho^k}\right)\left|\frac{f}{\rho^k}\right|^{q-1}x_i dx$$

$$=-\frac{q}{n}\sum_1^n\int\operatorname{sign}(f)\left(\frac{D_if}{\rho}-\frac{kfD_i\rho}{\rho^{k+1}}\right)\left|\frac{f}{\rho^k}\right|^{q-1}x_i dx.$$

Hence we have the identity

$$\text{(1)} \qquad \int\left[1-\frac{qk\sum D_i\rho x_i}{n\rho}\right]\left|\frac{f}{\rho^k}\right|^q dx=-\frac{q}{n}\sum\int\frac{\operatorname{sign}(f)x_i}{\rho}\frac{D_if}{\rho^{k-1}}\left|\frac{f}{\rho^k}\right|^{q-1}dx.$$

(i) In this case the left side is $\geqq \delta \int |f/\rho^k|^q dx$. Since $|\operatorname{sign}(f) x_i|/\rho \leqq 1$ the Hölder inequality for $q_1, q_2 > 1$, $1/q_1 + 1/q_2 = 1$ implies that

$$\int \left|\frac{f}{\rho^k}\right|^q dx \leqq \frac{q}{n\delta} \sum_1^n \left(\int \left| \frac{D_i f}{\rho^{k-1}} \right|^{q_1} dx \right)^{1/q_1} \cdot \left(\int \left| \frac{f}{\rho^k} \right|^{(q-1)q_2} dx \right)^{1/q_2}.$$

Choosing $q_1 = q$, $q_2 = q/(q-1)$ and dividing by $(\int |f/\rho^k|^q dx)^{(q-1)/q}$, we get the assertion.

(ii) In this case the left side is $\leqq -\delta \int |f/\rho^k|^q dx$. We divide by $-\delta$ and get

$$\int \left|\frac{f}{\rho^k}\right|^q dx \leqq \frac{q}{\delta n} \sum_1^n \int \frac{\operatorname{sign}(f) x_i}{\rho} \cdot \frac{D_i f}{\rho^{k-1}} \left| \frac{f}{\rho^k} \right|^{q-1} dx$$

$$\leqq \frac{q}{\delta n} \sum_1^n \int \left| \frac{D_i f}{\rho^{k-1}} \right| \cdot \left| \frac{f}{\rho^k} \right|^{q-1} dx.$$

Now the assertion follows as in case (i). Q.E.D.

Remark 2.2. Of course we did not need the assumption $k \in N$ in the proof of Lemma 2.1. Furthermore, it should be noted that in case $\rho(x)$ grows stronger than $O(|x|)$ the estimate $|\operatorname{sign}(f) x_i/\rho| \leqq 1$ is not sharp. This observation leads one to sharper estimates and yields essential improvements for second-order operators (see Janßen [6]).

THEOREM 2.3. *Let* Ω, p, k, n, δ *fulfill any of the following conditions*:

(i) $\rho(x) = (1 + |x|_1)^\alpha$, $\alpha \geqq 1$, $n - 2\alpha k > 0$, $\delta = (n - 2\alpha k)/n$.

(ii) $\rho(x) = (1 + |x|_1)^\alpha$, $\alpha \geqq 1$, $n - 2\alpha k < 0$, $\delta = (2\alpha k - n)/(n + 2\alpha k)$, $r = 2n/(2\alpha k - n)$, $\Omega \subset \{x \mid |x|_1 \geqq r\}$.

(iii) $2k < n$, $\rho(x) = (|x|_1 + c) \ln(|x|_1 + c)$, $c = \exp(4k/(n-2k))$, $\delta = (n-2k)/(2n)$.

(iv) $2k > n$, $\rho(x) = (|x|_1 + c) \ln(|x|_1 + c)$, $\delta = (2k-n)/(2n)$, *let* r, c *be chosen such that for* $|x|_1 \geqq r$ *we have* $|x|_1 \ln(|x|_1 + c) + |x|_1 \geqq ((2k+n)/4k)(|x|_1 + c) \ln(|x|_1 + c)$ *and* $\Omega \subset \{x \mid |x| \geqq r\}$.

(v) $\rho(x) = x_1$, $\Omega \subset R_+ \times R^{n-1}$, $\delta = (2k-1)/\sqrt{n}$.

(vi) $\Omega = R_+ \times R_+$, $\rho(x) = \operatorname{dist}(x, \partial\Omega)^\alpha = \min(x_1, x_2)^\alpha$, $2\alpha k \neq n = 2$, $\delta = (n - 2\alpha k)/n$.

Then for all $f \in \overset{0}{W}{}^{k,\rho}(\Omega)$

$$\int_\Omega \left(\frac{f}{\rho^k} \right)^2 dx \leqq \frac{4}{n\delta^2} \sum_1^n \int_\Omega \left(\frac{D_i f}{\rho^{k-1}} \right)^2 dx.$$

Proof. (i) $D_i \rho = \alpha (1 + |x|_1)^{\alpha - 1} \operatorname{sign} x_i$ implies that

$$1 - \frac{2k}{n\rho} \sum_1^n D_i \rho x_i = 1 - \frac{2\alpha k}{n} \frac{|x|_1}{1 + |x|_1} \geqq 1 - \frac{2\alpha k}{n} = \delta.$$

Now Lemma 2.1 implies that (after squaring both sides) for all $f \in C_0^\infty(\Omega)$

$$\int \left(\frac{f}{\rho^k} \right)^2 dx \leqq \frac{4}{n^2\delta^2} \left(\sum_1^n \left(\int \left(\frac{D_i f}{\rho^{k-1}} \right)^2 dx \right)^{1/2} \right)^2 \leqq \frac{4}{n\delta^2} \sum_1^n \int \left(\frac{D_i f}{\rho^{k-1}} \right)^2 dx.$$

Since $C_0^\infty(\Omega)$ lies dense in $\overset{0}{W}{}^{k,\rho}(\Omega)$, this gives our assertion.

(ii) As above we get

$$1 - \frac{2k}{n\rho} \sum_1^n D_i \rho x_i = 1 - \frac{2\alpha k}{n} \frac{|x|_1}{1 + |x|_1} \leqq 1 - \frac{2\alpha k}{n} \frac{r}{1 + r} = -\delta.$$

Now multiply by -1 and proceed as in (i).

(iii)

$$1-\frac{2k\sum D_i\rho x_i}{n\rho}=1-\frac{2k}{n}\,\frac{|x|_1(\ln(|x|_1+c)+1)}{(|x|_1+c)\ln(|x|_1+c)}$$

$$\geqq 1-\frac{2k}{n}\left(1+\frac{1}{\ln c}\right)=\delta.$$

(iv)

$$1-\frac{2k\sum D_i\rho x_i}{n\rho}\leqq 1-\frac{2k}{n}\,\frac{2k+n}{4k}=-\delta.$$

(v) Since

$$\int\left(\frac{f}{x_1^k}\right)^2 dx=-2\int D_1\left(\frac{f}{x_1^k}\right)\cdot\frac{f}{x_1^k}\cdot x_1\,dx$$

$$=-2\int\frac{D_1 f}{x_1^{k-1}}\cdot\frac{f}{x_1^k}dx+2k\int\left(\frac{f}{x_1^k}\right)^2 dx,$$

we get

$$\int\left(\frac{f}{x_1^k}\right)^2=\frac{2}{2k-1}\int\frac{D_1 f}{x_1^{k-1}}\,\frac{f}{x_1^k}\,dx\leqq\frac{2}{2k-1}\left(\int\left(\frac{D_1 f}{x_1^{k-1}}\right)^2 dx\right)^{1/2}\cdot\left(\int\left(\frac{f}{x_1^k}\right)^2 dx\right)^{1/2}.$$

We divide by $\|fx_1^{-k}\|_2$, square both sides and thus get the assertion.

(vi) We have

$$\rho(x)=\begin{cases}x_1^\alpha & \text{if } x_1\leqq x_2,\\ x_2^\alpha & \text{else,}\end{cases}$$

$$D_1\rho(x)=\begin{cases}\alpha x_1^{\alpha-1} & \text{if } x_1\leqq x_2,\\ 0 & \text{else,}\end{cases}\qquad D_2\rho(x)=\begin{cases}0 & \text{if } x_1\leqq x_2,\\ \alpha x_2^{\alpha-1} & \text{else.}\end{cases}$$

Now $\sum D_i\rho x_i=\alpha\rho$ and hence

$$1-\frac{2k}{n\rho}\sum D_i\rho x_i=1-\frac{2\alpha k}{n}=\delta.$$

Lemma 2.1 then implies the assertion as usual. Q.E.D.

Remark 2.4. (i) If the assertion of Theorem 2.3 holds for a weight function ρ and if $\tilde\rho$ is any other (not necessarily continuous) weight function such that $c_1\rho(x)\leqq\tilde\rho(x)\leqq c_2\rho(x)$ for all $x\in\Omega$ and some constants $c_1,c_2>0$, then an inequality of the same type as in Theorem 2.3 holds for $\tilde\rho$—only the constant $4n^{-1}\delta^{-2}$ has to be changed!

(ii) Mazja [11] studies Poincaré-type inequalities using capacity theory. But the constants appearing in his conditions are hard to compute or even prove finite, especially in the critical cases $n\leqq 2k$.

(iii) Using a suitable transformation of the coordinates, we can of course replace the condition on Ω in Theorem 2.3(ii) and (iv) by $\overline{\Omega}\neq R^n$.

The next examples show that the conditions on Ω cannot be relaxed, at least for $k=1$.

Example 2.5. (i) Let $n=1$ and ρ be a weight function which is less or equal than one on the interval $[-1,1]$. We consider the functions

$$f_m(x) := \begin{cases} m & \text{if } x\in[0,1], \\ m-(x-1) & \text{if } x\in(1,m+1), \\ 0 & \text{if } x\geqq m+1, \\ f_m(-x) & \text{if } x\leqq 0. \end{cases}$$

Then for all $m\in N$

$$\int_{-\infty}^{\infty} f_m^2\rho^{-2}\,dx \geqq \int_{-1}^{1} f_m^2\,dx = 2m^2 \quad \text{and} \quad \int_{-\infty}^{\infty} (f_m')^2\,dx \leqq 2m.$$

Hence there cannot exist any locally bounded weight functions ρ and any constant K such that for all $f\in C_0^\infty(R)$

$$\int_{-\infty}^{\infty} f^2\rho^{-2}\,dx \leqq K\int_{-\infty}^{\infty} (f')^2\,dx.$$

(ii) Now let $n=2$ and ρ be a weight function which is smaller than or equal to one on the unit ball $K(0,1)$. For $m\in N$, $r\in R_+$ we define

$$g_m(r) := \begin{cases} m & \text{if } r\in[0,1], \\ m-\ln r & \text{if } r\in(1,\exp m), \\ 0 & \text{if } r\geqq \exp m. \end{cases}$$

Then

$$\int_{R^2} g_m(|x|)^2\rho^{-2}(x)\,dx \geqq \pi m^2 \quad \text{and} \quad \sum_1^2 \int_{R^2} \big(D_i g_m(|x|)\big)^2\,dx = w\pi m$$

and again there cannot exist any locally bounded weight function ρ and any constant K such that for all $f\in C_0^\infty(R^2)$

$$\int_{R^2} f^2\rho^{-2}\,dx \leqq K\sum_1^2 \int_{R^2} (D_i f)^2\,dx.$$

COROLLARY 2.6. *Let $\Omega\subset R^n$, ρ be a weight function, $k_0\in N$. If Ω, p, k, n fulfill for any $k\leqq k_0$ any of the conditions of Theorem* 2.3, *then there exists a constant K, which can be computed explicitly, such that for all $f\in \overset{0}{W}{}^{k,\rho}(\Omega)$*

$$\sum_{|\alpha|=k_0} \|D^\alpha f\|_2^2 \leqq \sum_{|\alpha|\leqq k_0} \|\rho^{-k_0+|\alpha|}D^\alpha f\|_2^2 \leqq K \sum_{|\alpha|=k_0} \|D^\alpha f\|_2^2.$$

Proof. The proof follows immediately from Theorem 2.3 by induction. Q.E.D.

We shall now apply our theorems to the Dirichlet problem on unbounded domains. The proof follows the philosophy of J. Wloka (see [18]).

THEOREM 2.7. *Let Ω, ρ, k, n fulfill the assumptions of Corollary* 2.6. *Furthermore, let $Au := \sum_{|\alpha|,|\beta|\leqq k}(-1)^{|\alpha|}D^\alpha(a_{\alpha\beta}D^\beta u)$, where the matrix $(a_{\alpha\beta})_{|\alpha|=|\beta|=k}$ is assumed to be uniformly positive definite and*

(i) $a_{\alpha\beta}\in L_\infty$ *for* $|\alpha|=|\beta|=k$,

(ii) $a_{\alpha\beta}=0$ *for* $k<|\alpha|+|\beta|<2k$,

(iii) $D^\gamma(a_{\alpha\beta})\rho^{2k-|\beta+\alpha-\gamma|}\in L_\infty$ *for* $|\alpha|+|\beta|\leqq k$, $\gamma\leqq\alpha$.

Set

$$a(u,v) := \sum_{|\alpha|,|\beta| \leqq k} \int_\Omega a_{\alpha\beta} D^\beta u D^\alpha v \, dx.$$

Then there exists a $\lambda_0 > 0$ *such that for all* $\lambda > \lambda_0$ *and all* $f \in (\overset{0}{W}{}^{k,\rho}(\Omega))^*$ *there is exactly one* $u \in \overset{0}{W}{}^{k,\rho}(\Omega)$ *such that for all* $v \in \overset{0}{W}{}^{k,\rho}(\Omega)$

$$b(u,v) := a(u,v) + \lambda \int_\Omega uv\rho^{-2k} dx = \langle f, v \rangle_{\overset{0}{W}{}^{k,\rho}}.$$

The operator $A + \lambda\rho^{-2k}: \overset{0}{W}{}^{k,\rho} \to (\overset{0}{W}{}^{k,\rho})^*$ *is a topological isomorphism.*

Proof. b is for all $\lambda \in R$ a continuous bilinear form. From the assumptions on the coefficients and Corollary 2.6 we conclude that

$$(2) \qquad a(u,u) = \sum_{|\alpha|=|\beta|=k} \int a_{\alpha\beta} D^\beta u D^\alpha u \, dx + \sum_{|\alpha|+|\beta|<k} \int a_{\alpha\beta} D^\beta u D^\alpha u \, dx$$

$$\geqq c_1 \sum_{|\alpha|=k} \|D^\alpha u\|_2^2 - c_2 \sum_{|\alpha|+|\beta|<k} \sum_{\gamma<\alpha} \int \left| D^\gamma(a_{\alpha\beta}) D^{\beta+\alpha-\gamma} u \cdot u \right| dx$$

$$\geqq c_3 \|u\|_{k,\rho}^2 - c_4 \|u\|_{k,\rho} \cdot \|u\rho^{-k}\|_2,$$

where c_i are positive constants. Now let $c_3 > \delta > 0$ and $\varepsilon := 2(c_3 - \delta)c_4^{-1}$; then

$$(3) \qquad c_4 \|u\|_{k,\rho} \|u\rho^{-k}\|_2 \leqq c_4 \varepsilon 2^{-1} \|u\|_{k,\rho}^2 + c_4 (2\varepsilon)^{-1} \|u\rho^{-k}\|_2^2$$

$$= (c_3 - \delta)\|u\|_{k,\rho}^2 + c_4^2 (4(c_3 - \delta))^{-1} \|u\rho^{-k}\|_2^2.$$

For any $\lambda > \lambda_0 := c_4^2 (4c_3)^{-1}$ exists a $\delta > 0$, such that $\lambda \geqq c_4^2 (4(c_3 - \delta))^{-1}$. For this δ it follows from (2), (3) that $b(u,u) \geqq \delta \|u\|_{k,\rho}^2$, i.e. b is coercive. The Lax–Milgram lemma (see e.g. Oden, Reddy [13]) gives us then the assertion. Q.E.D.

Remark 2.8. (i) λ_0 can be computed explicitly using Corollary 2.6 and bounds on the coefficients.

(ii) The conditions on $a_{\alpha\beta}$ for $|\alpha| + |\beta| < 2k$ will be weakened considerably in the next section. But then we use compactness arguments and shall not be able to compute λ_0 any more.

(iii) We can also treat mixed boundary conditions in this way. For an example see Janßen [5].

3. Compact imbeddings. In this section we study compact imbeddings and some of their consequences. We shall need some geometric properties of Ω and the weight function ρ, which we shall provide in the next definitions and lemmata.

DEFINITION 3.1. Let $l \in N$, $0 \leqq \kappa \leqq 1$ (for $l = 0$ only $\kappa = 0$ or 1).

(i) The function $f:\Omega\to R$ belongs to $C^{(l,\kappa)}(\Omega)$, if the numbers

$$\sup_{x\in\Omega}|D^\alpha f(x)|, \qquad |\alpha|\leqq l,$$

$$\sup_{x,y\in\Omega}\frac{|D^\alpha f(x)-D^\alpha f(y)|}{|x-y|}, \qquad |\alpha|\leqq l-1,$$

$$\sup_{x,y\in\Omega}\frac{|D^\alpha f(x)-D^\alpha f(y)|}{|x-y|^\kappa}, \qquad |\alpha|=l.$$

are all finite.

(ii) A one-to-one mapping $T:\Omega\to\Omega'\subset R^n$ is called a $C^{(l,\kappa)}$-diffeomorphism if the components of T resp. T^{-1} belong to $C^{(l,\kappa)}(\Omega)$ resp. $C^{(l,\kappa)}(\Omega')$ and for $l\geqq 1$ $0<c_1\leqq|\det(\partial T/\partial x)(x)|\leqq c_2<\infty$ for all $x\in\Omega$.

(iii) $\Omega\subset R^n$ is called (l,κ)-smooth if for all $x\in\partial\Omega$ there exists a neighbourhood U_x such that U_x is (l,κ)-diffeomorphic to the unit cube W at which (all mappings are one-to-one, onto)

$$\partial\Omega\cap U_x\to W\cap\{x_n=0\},$$
$$\Omega\cap U_x\to W\cap\{x_n>0\},$$
$$C\bar{\Omega}\cap U_x\to W\cap\{x_n<0\}.$$

LEMMA 3.2. *Let $T:\Omega\to\Omega'$ be a $C^{(L,\kappa)}$-diffeomorphism. Then the pullback operators*

$$*T: W^{l,\rho}(\Omega')\to W^{l,\rho\circ T}(\Omega), \qquad u\to u\circ T$$

$$*T^{-1}: W^{l,\rho\circ T}(\Omega)\to W^{l,\rho}(\Omega'), \qquad u\to u\circ T^{-1}$$

are continuous for $l\leqq L+\kappa$ and hence

$$W^{l,\rho}(\Omega')\cong W^{l,\rho\circ T}(\Omega).$$

Proof. Analogous to the case $\rho\equiv 1$, see e.g. Wloka [18, p. 86ff], or Adams [1]. Q.E.D.

DEFINITION 3.3. A weight function $\rho:R^n\to R_+$ is called translation, and dilation invariant (td-function) if for all $K_0>0$ and all $x_0\in R^n$ there exists a constant K_1 such that for the mapping $T:R^n\to R^n$, $T(x)=K_0(x-x_0)$ and all $x\in R^n$ $K_1^{-1}\rho(Tx)\leqq\rho(x)\leqq K_1\rho(Tx)$, where K_1 as a function of K_0, x_0 is locally bounded.

LEMMA 3.4. (i) $1+|x|$ *and* $\ln(|x|+2)$ *are td-functions.*

(ii) *If ρ_1, ρ_2 are td-functions, then the same is true for $\rho_1\cdot\rho_2$, ρ_1^c for any $c>0$ and also any weight-function ρ with $c_1^{-1}\rho_1<\rho<c_1\rho_1$, $c_1>0$ is a td-function.*

(iii) *If ρ is a td-function and $f:R^+\to R^+$ increasing, then $f\circ\rho$ is a td-function.*

Proof. Left to the reader.

LEMMA 3.5. *Let $M\in R_+$, $K_0\in R_+$ and $x_0\in R^n$ such that $K_0+|x_0|\leqq M$. If ρ is a td-function, $T(x):=K_0(x-x_0)$ and $\Omega':=T\Omega$, then for all $l\in N$, $W^{l,\rho}(\Omega)\cong W^{l,\rho}(\Omega')$. The norms of the pullback operators depend only on l and M.*

Proof. T is a C^∞-diffeomorphism. Lemmata 3.2 and 3.4(ii) imply $W^{l,\rho}(\Omega)\cong W^{l,\rho}(\Omega')$ and the continuity of the pullback operators. Since by Definition 3.3 K_1 as a function of K_0, x_0 is locally bounded, the norms depend only on M (and of course 1). Q.E.D.

LEMMA 3.6. *Let $T:\Omega\to\Omega'$ be a $C^{(L,\kappa)}$-diffeomorphism, $L+\kappa\geqq 1$. If $\rho(x)=f(1+|x|)$, $f:R_+\to R_+$ increasing, then for all $l\leqq L+\kappa$, $W^{l,\rho}(\Omega)\cong W^{l,\rho}(\Omega')$.*

Proof. ρ is a td-function; hence by 3.5, without loss of generality, $0\in\Omega\cap\Omega'$, $T(0)=0$. T, T^{-1} are Lipschitz continuous; hence there exists a $K_1>0$ such that $K_1^{-1}|x|\leqq|Tx|\leqq K_1|x|$. Since f is increasing, it follows that $\rho(K_1^{-1}x)\leqq\rho(Tx)\leqq\rho(K_1x)$. Since ρ is dilation invariant, there exists $K_2>0$ such that $K_2^{-1}\rho(x)\leqq\rho(Tx)\leqq K_2\rho(x)$ and Lemma 3.2 gives us the assertion. Q.E.D.

Before we proceed, let us fix some more notation. If $k\in N$, ρ is a weight function on Ω, then $M_{k,\rho}$ denotes the operator which assigns to any measurable $f:\Omega\to R$ the function $M_{k,\rho}(f):=\rho^{-k}\cdot f$. If $\phi:\Omega\to R$ is measurable, then $M_\phi(f):=\phi\cdot f$. Let $\Sigma\subset\{x\in R^n\,|\,|x|=1\}$, $x_0\in R^n$, $r_0>0$, then

$$S(x_0,r_0,\Sigma):=\{x_0+t\cdot\sigma\,|\,r_0\leqq t,\ \sigma\in\Sigma\}.$$

For $a\geqq 1$ we introduce the following abbreviation, where $\rho(x)=\tilde\rho(|x|)$ (for shortness $\rho(|x|)$)

$$I(a,n,k,\rho):=\int_a^\infty\frac{\int_1^u v^{1-n}\rho^{2(k-1)}(v)\,dv}{\rho^{2k}(u)}\cdot u^{n-1}\,du.$$

Furthermore, we shall need the following well-known theorem due to Kolmogoroff (for a proof see Yoshida [19] or Voigt and Wloka [17]):

$K\subset L_2(\Omega)$ is relatively compact iff the following three conditions hold:

(K1) K is bounded in $L_2(\Omega)$;

(K2) $\lim_{h\to 0}\int_\Omega|f(x+h)-f(x)|^2dx=0$ uniformly for all $f\in K$;

(K3) $\lim_{a\uparrow\infty}\int_{|x|>a}|f(x)|^2dx=0$ uniformly for all $f\in K$.

LEMMA 3.7. *Let $\Omega=S(0,1,\Sigma)$, $k\in N$, ρ be a weight function with $\rho(x):=\tilde\rho(|x|)$ (for shortness $\rho(|x|)$) and $f\in C^\infty(\Omega)$ with $\|f\|_{k,\rho}\leqq 1$ and $f(\sigma)=0$ for $\sigma\in\Sigma$. Then*

$$\int_{\Omega\cap\{|x|>a\}}f^2(x)\rho^{-2k}(x)\,dx\leqq I(a,n,k,\rho).$$

Proof. We introduce polar coordinates $x=T(u,\alpha_1,\cdots,\alpha_{n-1})$.

$$\begin{aligned}
x_1&=u\cos\alpha_1,\\
x_2&=u\sin\alpha_1\cos\alpha_2,\\
&\vdots\\
x_{n-1}&=u\sin\alpha_1\cdots\sin\alpha_{n-2}\cos\alpha_{n-1},\\
x_n&=u\sin\alpha_1\cdots\sin\alpha_{n-2}\sin\alpha_{n-1},
\end{aligned}$$

where $0\leqq u<\infty$, $0\leqq\alpha_{n-1}\leqq 2\pi$, $0\leqq\alpha_i\leqq\pi$ for $i=1,\cdots,n-2$ and set $\tilde f:=f\circ T$. Since $\tilde f(1,\alpha_1,\cdots,\alpha_{n-1})=0$ for $(\alpha_1,\cdots,\alpha_{n-1})\in T^{-1}(1,\Sigma)=:\tilde\Sigma$ we have for $x=T(u,\alpha_1,\cdots,\alpha_{n-1})\in\Omega$ that

$$\tilde f(u,\alpha_1,\cdots,\alpha_{n-1})=\int_1^u\frac{\partial\tilde f}{\partial v}(v,\alpha_1,\cdots,\alpha_{n-1})\,dv.$$

This implies that

$$\begin{aligned}
|\tilde f(u,\alpha_1,\cdots,\alpha_{n-1})|^2&=\left|\int_1^u\frac{\partial\tilde f}{\partial v}(v,\alpha_1,\cdots,\alpha_{n-1})\,dv\right|^2\\
&=\left(\int_1^u\frac{|(\partial\tilde f/\partial v)(v,\alpha_1,\cdots,\alpha_{n-1})|^2}{\rho^{2(k-1)}(v)}\cdot v^{(n-1)/2}\cdot\rho^{(k-1)}(v)\cdot v^{(1-n)/2}\,dv\right)^2\\
&\leqq F(u,n,k,\rho)\int_1^\infty\frac{|(\partial\tilde f/\partial v)(v,\alpha_1,\cdots,\alpha_{n-1})|^2}{\rho^{2(k-1)}(v)}\cdot v^{n-1}\,dv
\end{aligned}$$

where $F(u,\ n,\ k,\ \rho) := \int_1^u v^{1-n}\rho^{2(k-1)}(v)\,dv$. We multiply both sides of the inequality by $\Pi_1^{n-2}(\sin\alpha_i)^{n-i-1}$, integrate with respect to $\alpha_1,\cdots,\alpha_{n-1}$ over $\tilde{\Sigma}$ and since

$$\left|\frac{\partial\tilde{f}}{\partial v}(y)\right|^2 \leqq \left(\sum_1^n \left|\frac{\partial f}{\partial x_i}(Ty)\right|^2\right)\left(\sum_1^n \left|\frac{\partial T_i}{\partial v}(y)\right|^2\right) \leqq \sum \left|\frac{\partial f}{\partial x_i}(Ty)\right|^2$$

and $\|f\|_{k,\rho}\leqq 1$ we arrive after a coordinate transformation on the right side at

$$\int\cdots\int_{\tilde{\Sigma}} |\tilde{f}(u,\alpha_1,\cdots,\alpha_{n-1})|^2 \prod_1^{n-2}(\sin\alpha_i)^{n-i-1}\,d\alpha_1\cdots d\alpha_{n-1} \leqq F(u,\ n,\ k,\ \rho).$$

Multiplying by $u^{n-1}\rho^{-2k}(u)$, integrating with respect to u from a to ∞ and transforming the coordinates on the left side, we finally get the assertion. Q.E.D.

LEMMA 3.8. *If ρ, k, n fulfill any of the following conditions*

(i) $2k>n$, $\rho(x)=(|x|+2)\ln(|x|+2)$,

(ii) $\rho(x)=(|x|+1)^c$, $c>\max(1,n/2k)$,

(iii) $\rho(x)=(|x|+2)^c\cdot\ln(|x|+2)$, $c>1$, $n=2kc$,

then ρ is a td-function, $\rho^{-k}\in L_2(R^n)$ and $\lim_{a\to\infty} I(a,n,k,\rho)=0$.

Proof. Lemma 3.4 implies that all functions are td-functions. In all three cases we easily see that $\int_0^\infty \rho^{-2k}(u)u^{n-1}\,du<\infty$ and hence $\rho^{-k}\in L_2(R^n)$.

(i) Set $F(u,n,k,\rho)=\int_1^u v^{1-n}\rho^{2(k-1)}(v)\,dv$; then

$$F(u,n,k,\rho)\leqq c_1\int_1^u v^{1-n+2(k-1)}(\ln(v+2))^{2(k-1)}\,dv.$$

Since $2k>n$ we have $m := 1-n+2(k-1)\geqq 0$ and hence

$$\frac{F(u,n,k,\rho)}{\rho^{2k}(u)} \leqq c_1\frac{\int_1^u v^m\,dv}{u^{2k}\ln^2(u+2)} \leqq c_2\frac{u^{m+1-2k}}{\ln^2(u+2)} = c_2\frac{u^{-n}}{\ln^2(u+2)}.$$

For $a\geqq 2$ this implies that

$$I(a,n,k,\rho)\leqq c_3\int_a^\infty \frac{1}{u\ln^2 u}\,du = \frac{c_3}{\ln a} \overset{a\to\infty}{\to} 0.$$

(ii) Now we define $m := 1-n+2(k-1)c$. Then

$$F(u,n,k,\rho)\leqq c_4\int_1^u v^m\,dv.$$

Case 1. $m<-1$. Then $F(u,n,k,\rho)\leqq c_4$ and hence $I(a,n,k,\rho)\leqq c_4\int_a^\infty u^{n-2kc-1}\,du\to 0$ for $a\to\infty$ since $n-2kc-1<-1$.

Case 2. $m=-1$. Then $F(u,n,k,\rho)\leqq c_4\ln u$ and hence $I(a,n,k,\rho)\leqq c_4\int_a^\infty \ln(u)\cdot u^{n-2kc-1}\,du\leqq c_5\int_a^\infty u^{-1-\delta}\,du\to 0$ for $a\to\infty$ where we choose $\delta>0$ such that $\ln u\leqq c_5u^\delta/c_4$ and $n-2kc-1\leqq -1-2\delta$.

Case 3. $m>-1$. Then $F(u,n,k,\rho)\leqq c_6u^{m+1}$ and $I(a,n,k,\rho)\leqq c_6\int_a^\infty u^{-2c+1}\,du\to 0$ for $a\to\infty$, since $-2c+1<-1$.

(iii) For $a\geqq 2$ we get

$$I(a,n,k,\rho)\leqq c_7\int_a^\infty \frac{\int_1^\infty v^{1-2c}\,dv}{u^{2kc}\ln^2 u}\cdot u^{n-1}\,du \leqq c_7\int_a^\infty \frac{1}{u\ln^2 u}\,du = \frac{c_7}{\ln a} \overset{a\to\infty}{\to} 0.$$ Q.E.D.

THEOREM 3.9. *Let ρ, k, n fulfill any of the conditions of Lemma 3.8. Then $M_{k,\rho}: W^{k,\rho}(R^n)\to L_2(R^n)$ is a linear, continuous and injective operator with dense range which maps bounded into relatively compact sets. (Henceforth we shall call a mapping with these properties a compact imbedding.)*

Proof. Obviously, $M_{k,\rho}$ is linear, continuous, injective and has dense image. Let $\phi_1 \in C_0^\infty(R^n)$ with $\operatorname{supp}\phi_1 \subset K(0,2)$ and $\phi_1(x)=1$ for $x \in K(0,1)$. Since

$$M_{\phi_1}: W^{k,\rho}(R^n) \to \overset{0}{W}{}^{k,\rho}(K(0,2)) \cong \overset{0}{W}{}^{k}(K(0,2))$$

is continuous, we get from the imbedding theorems on bounded domains that

$$M_{k,\rho} \circ M_{\phi_1}: W^{k,\rho}(R^n) \to L_2(K(0,2)) \subset L_2(R^n)$$

is compact. If we define $\phi_2 := 1-\phi_1$, then the derivatives of ϕ_2 have compact support and since the weight functions are bounded from above and below by positive constants, it follows that

$$M_{\phi_2}: W^{k,\rho}(R^n) \to \overset{0}{W}{}^{k,\rho}(R^n \backslash K(0,1))$$

is continuous. Since $M_{k,\rho} = M_{k,\rho} \circ M_{\phi_1} + M_{k,\rho} \circ M_{\phi_2}$, it suffices to show that

$$M_{k,\rho}\big(M_{\phi_2}(K)\big) \subset L_2(R^n \backslash K(0,1))$$

is relatively compact where $K := \{f \in W^{k,\rho} \mid \|f\|_{k,\rho} \leqq 1\}$.

Now (K1) is immediate and (K3) follows from Lemmata 3.7, 3.8. Since M_{ϕ_2} is continuous, there exists $c_1 > 0$ such that

$$M_{\phi_2}(K) \subset \overline{\{f \in C^\infty(R^n) \mid \|f\|_{k,\rho} \leqq c_1,\ \operatorname{supp} f \subset R^n \backslash K(0,1)\}}^{\|\cdot\|_{k,\rho}}.$$

So let $f \in C^\infty(R^n)$, $\|f\|_{k,\rho} \leqq c_1$, $\operatorname{supp} f \subset R^n \backslash K(0,1)$. Then

$$\begin{aligned}
&\int |M_{k,\rho}(f)(x+h) - M_{k,\rho}(f)(x)|^2\, dx \\
&= \int \left| \frac{f(x+h)}{\rho^k(x+h)} - \frac{f(x)}{\rho^k(x)} \right|^2 dx \\
&= \int \left| \frac{f(x+h)\rho^k(x) - f(x)\rho^k(x) + f(x)\rho^k(x) - f(x)\rho^k(x+h)}{\rho^k(x+h)\rho^k(x)} \right|^2 dx \\
&\leqq 2\int \left| \frac{f(x+h)-f(x)}{\rho^k(x+h)} \right|^2 + \left(\frac{f(x)}{\rho^k(x)} \right)^2 \left| \frac{\rho^k(x)-\rho^k(x+h)}{\rho^k(x+h)} \right|^2 dx \\
&\leqq 2\int \left\{ \frac{\left(\int_0^1 \langle \nabla f(x+th) \mid h \rangle\, dt \right)^2}{\rho^{2k}(x+h)} + \left(\frac{f(x)}{\rho^k(x)} \right)^2 \frac{\left(\int_0^1 \langle \nabla \rho^k(x+th) \mid h \rangle\, dt \right)^2}{\rho^{2k}(x+h)} \right\} dx \\
&\leqq 2|h|^2 c_2 \int_0^1 \int \left\{ \frac{|\nabla f(x+th)|^2}{\rho^{2(k-1)}(x+th)} + \left(\frac{f(x)}{\rho^k(x)} \right)^2 \right\} dx\, dt \qquad \text{(since } \rho \text{ is a td-function)} \\
&\leqq c_3 |h|^2 \to 0 \quad \text{for } |h| \to 0
\end{aligned}$$

where c_3 does not depend on f (only on c_1, k, ρ). This implies (K2) and hence the assertion is proved. Q.E.D.

COROLLARY 3.10. *Let $\Omega \subset R^n$, k, n, ρ as in Theorem 3.9. Then $M_{k,\rho}$: $\overset{0}{W}{}^{k,\rho}(\Omega) \to L_2(\Omega)$ is a compact imbedding.*

Proof. Let $F_\Omega: \overset{0}{W}{}^{k,\rho}(\Omega) \to \overset{0}{W}{}^{k,\rho}(R^n)$ be the natural continuation operator (zero outside Ω) and $R_\Omega: L_2(R^n) \to L_2(\Omega)$ the restriction mapping. Then F_Ω, R_Ω are continuous and hence

$$M_{k,\rho}^{\Omega} = R_\Omega \circ M_{k,\rho}^{R^n} \circ F_\Omega$$

is compact by Theorem 3.9. Q.E.D.

If there is a continuation operator for the spaces $W^{k,\rho}$, i.e. a continuous, linear mapping $F_\Omega: W^{k,\rho}(\Omega) \to W^{k,\rho}(R^n)$ with $Fu(x) = u(x)$ for all $x \in \Omega$, $u \in C^k(\Omega) \cap W^{k,\rho}(\Omega)$, we can easily transfer Corollary 3.10 to these spaces. We now give some examples where such an operator exists.

LEMMA 3.11. *Let* $\Omega = \{x \in R^n \mid x_n > 0\}$. *Then there exists a continuation operator.*

Proof. Use the method of Hestenes, see Adams [1], Wloka [18]. Q.E.D.

Remark 3.12. By multiple application of Hestenes' construction we can of course also construct continuation operators for orthants and similar domains, e.g. $\{x \in R^3 \mid x_1 > 0 \text{ and } x_2 > 0\}$.

LEMMA 3.13. (i) *Let* Ω *be* $C^{(k,\kappa)}$*-diffeomorphic to a domain* Ω' *for which there exists a continuation operator for* $l \leqq k + \kappa$. *Then there exists also a continuation operator for* Ω *and* $l \leqq k + \kappa$.

(ii) *Let* Ω *be* (k, κ)*-smooth with bounded boundary. Then there exists a continuation operator.*

Proof. (i) $F_\Omega := {}^*T \circ F_{\Omega'} \circ {}^*T^{-1}$ (see Lemma 3.2).

(ii) Cover $\partial\Omega$ by small neighbourhoods U_j, $j = 1, \cdots, m$ from Lemma 3.2(iii). Then (with $U_0 = \Omega$) we have $\overline{\Omega} \subset \bigcup_{j=0}^{m} U_j$. Let α_j, $j = 0, \cdots, m$ be the appropriate partition of the unity. Then α_0 is equal to one outside $\bigcup_1^m U_j$; hence all derivations of α_0 (and of course those of α_j, $j = 1, \cdots, m$) vanish outside a bounded ball and hence $|D^\beta \alpha_j \circ \rho^l|$ stays uniformly bounded. This implies that $M_{\alpha_j}: W^{l,\rho}(U_j \cap \Omega) \to W^{l,\rho}(\Omega)$ is continuous and their norm is uniformly bounded. By (i) we can extend $\alpha_j f$ for $j = 1, \cdots, m$ to $U_j (f \in W^{l,\rho}(\Omega))$. Adding these local extensions to $\alpha_0 f$, we get an extension with support in $\bigcup_{j=0}^{m} U_j$ and can extend by zero to the whole space. Q.E.D.

THEOREM 3.14. *Let* $\Omega = \bigcup_1^N \Omega_i$, *where there exists a continuation operator for each* Ω_i *and* ρ, k, n *fulfill the conditions of Lemma* 3.8. *Then* $M_{k,\rho}: W^{k,\rho}(\Omega) \to L_2(\Omega)$ *is a compact imbedding.*

Proof. For $N = 1$, analogous to Corollary 3.10. For $N > 1$ let $(\phi_j)_{j \in N} \subset W_{k,\rho}(\Omega)$ with $\|\phi_j\|_{k,\rho} \leqq 1$. Then $\phi_j \in W^{k,\rho}(\Omega_i)$, $i = 1, \cdots, N$ and there exist (eventually after passing to some convergent subsequence several times) $\psi_i \in L_2(\Omega_i)$ such that $M_{k,\rho}\phi_j \to \psi_i$ in $L_2(\Omega_i)$. Now define

$$\psi := \psi_i \quad \text{on } \Omega_i \setminus \bigcup_{l=1}^{i-1} \Omega_l, \qquad i = 1, \cdots, N.$$

Then $M_{k,\rho}\phi_j \to \psi$ in $L_2(\Omega)$ and $M_{k,\rho}$ is compact. Q.E.D.

For $n > 2k$ the conditions on the weight function are very restrictive (see Lemma 3.8). This defect can be removed if we add a decay condition at infinity, i.e. we consider now the spaces $W_c^{k,\rho}$.

LEMMA 3.15. *Let* $n > 2k$ *and* $c > 1$ *be such that* $n > 2kc + 2(1 - c)$. *If* $\Omega = S(0, 1, \Sigma)$ *and* $\rho(x) = (1 + |x|)^c$ *then*

$$\int_{\Omega \cap \{|x| \geqq a\}} \left(\frac{f(x)}{\rho^k(x)} \right)^2 dx \to 0 \quad \text{as } a \to \infty$$

uniformly for all $f \in W_c^{k,\rho}(\Omega)$ *with* $\|f\|_{k,\rho} \leqq 1$.

Proof. Let $f \in C_0^\infty(R^n)|_\Omega$, introduce polar coordinates $x = T(u, \alpha_1, \cdots, \alpha_{n-1})$ and set $\tilde{f} := f \circ T$. Then

$$\tilde{f}(u, \alpha_1, \cdots, \alpha_{n-1}) = \int_u^\infty \frac{\partial \tilde{f}}{\partial v}(v, \alpha_1, \cdots, \alpha_{n-1})\, dv.$$

Similar to Lemma 3.7 this implies that

$$\int_{\Omega \cap \{|x| \geqq a\}} \frac{f^2(x)}{\rho^{2k}(x)}\, dx \leqq \int_a^\infty \frac{G(u,n,k,\rho)}{\rho^{2k}(u)} u^{n-1}\, du$$

where $G(u,n,k,\rho) := \int_u^\infty v^{1-n} \rho^{2(k-1)}(v)\, dv$. Since

$$\int_a^\infty \frac{G(u,n,k,\rho)}{\rho^{2k}(u)} u^{n-1}\, du \leqq c_1 a^{2-2c}$$

the assertion follows from $c > 1$. Q.E.D.

THEOREM 3.16. *Let $n > 2k$, $d > 1$, $n > 2kd + 2(1-d)$, $\rho(x) = (1 + |x|)$. Then*

(i) *$M_{k,\rho}: W_c^{k,\rho}(R^n) \to L_2(R^n)$ is a compact imbedding.*

(ii) *If $\Omega = \bigcup_1^N \Omega_i$ and for each Ω_i exists a continuation operator, then $M_{k,\rho}: W_c^{k,\rho}(\Omega) \to L_2(\Omega)$ is a compact imbedding.*

(iii) *For any $\Omega \subset R^n$ $M_{k,\rho}: W_c^{k,\rho}(\Omega) \to L_2(\Omega)$ is a compact imbedding.*

Proof. Analogous to that of Theorem 3.9, Corollary 3.10 and Theorem 3.14, using Lemma 3.15 instead of Lemmata 3.7, 3.8. Q.E.D.

THEOREM 3.17. *Let $M_{j,\rho}: W^{j,\rho}(\Omega) \to L_2(\Omega)$ be compact for $1 \leqq j \leqq k$. Then for any $\varepsilon > 0$ there exists a $c(\varepsilon) > 0$ such that for all $f \in W^{k,\rho}(\Omega)$*

$$\sum_{|\alpha| \leqq k-1} \| D^\alpha f \rho^{-k+|\alpha|} \|_2 \leqq \varepsilon \| f \|_{k,\rho} + c(\varepsilon) \| f \rho^{-k} \|_2$$

(*similar for* $\overset{0}{W}{}^{k,\rho}$, $W_c^{k,\rho}$).

Proof. Suppose the assertion wrong. Then there exist $\varepsilon > 0$ and a sequence $f_m \in W^{k,\rho}$ with $\|f_m\|_{k,\rho} = 1$ such that for all $m \in N$

$$(*) \qquad \sum_{|\alpha| \leqq k-1} \| D^\alpha f_m \rho^{-k+|\alpha|} \|_2 > \varepsilon + m \| f_m \rho^{-k} \|_2.$$

From the compactness it follows that there exists a subsequence (again denoted by f_m) and $g_\alpha \in L_2$ such that $D^\alpha f_m \rho^{-k+|\alpha|} \to g_\alpha$ for $|\alpha| \leqq k-1$. We define $h_\alpha := g_\alpha \rho^{k+|\alpha|}$. Then $D_\alpha h_0 = h_\alpha$ in the sense of distributions. Since $\|f_m\|_{k,\rho} \leqq 1$ $(*)$ implies that $\|f_m \rho^{-k}\| \to 0$ and $g_0 = 0$. Hence h_0 and also $h_\alpha = 0$ for $|\alpha| \leqq k-1$. Going to the limit in $(*)$ this implies that $0 > \varepsilon$ and this is the desired contradiction. Q.E.D.

Remark 3.18. (i) Theorem 3.17 is a generalization of Ehrling's lemma (see Wloka [18]).

(ii) Benci and Fortunato [2] prove an imbedding theorem of the following kind: Let $\phi := \psi \cdot \rho^{-k}$, $\psi(x) \to 0$ for $|x| \to \infty$ then $M_\phi: \overset{0}{W}{}^{k,\rho} \to L_2$ is compact. This result can also be applied to study Fredholm properties of elliptic operators but from it we cannot derive Ehrling's lemma or variants of Poincaré's lemma for functions not vanishing on the boundary (see §4).

(iii) Imbedding theorems for weighted Sobolev spaces have also been studied by Otelbaev [14]. Contrary to our result, there ρ^{-k} has to grow sufficiently fast.

As an application of these imbedding theorems we study again the Dirichlet problem.

THEOREM 3.19. *Let* Ω, ρ, k, n *fulfill the conditions of Corollary* 2.6 *and Theorem* 3.17. *Furthermore, let* $Au := \sum_{|\alpha|,|\beta| \leq k}(-1)^{|\alpha|} D^{\alpha}(a_{\alpha\beta} D^{\beta} u)$ *where* $(a_{\alpha\beta})_{|\alpha|=|\beta|=k}$ *is uniformly strict positive definite and* $a_{\alpha\beta}\rho^{2k-|\alpha|-|\beta|} \in L_{\infty}$. *Again let* $a(u,v) := \sum \int a_{\alpha\beta} D^{\beta} u D^{\alpha} v \, dx$; *then it follows that*

(i) *There exists* λ_0 *such that for all* $f \in (\overset{0}{W}{}^{k,\rho}(\Omega))^*$ *and for all* $\lambda > \lambda_0$ *exists one and only one* $u \in \overset{0}{W}{}^{k,\rho}(\Omega)$ *such that for all* $v \in \overset{0}{W}{}^{k,\rho}(\Omega)$

$$b_{\lambda}(u,v) := a(u,v) + \lambda \int uv\rho^{-2k} dx = \langle f, v \rangle_{k,\rho}.$$

(ii) $A + \lambda\rho^{-2k} : \overset{0}{W}{}^{k,\rho}(\Omega) \to (\overset{0}{W}{}^{k,\rho}(\Omega))^*$ *is an isomorphism for all* $\lambda > \lambda_0$.

(iii) $A : \overset{0}{W}{}^{k,\rho}(\Omega) \to (\overset{0}{W}{}^{k,\rho}(\Omega))^*$ *is a Fredholm operator with index zero.*

(iv) *There exist only countably many* λ *such that* $Av + \lambda\rho^{-2k} v = 0$ *admits a nontrivial solution. These eigenvalues have no finite accumulation point.*

Proof. b is continuous and from Corollary 2.6 and Theorem 3.17 we get

$$a(u,u) \geqq c_1 \| u \|^2_{k,\rho} - c_2 \| u\rho^{-k} \|_2, \qquad c_1 > 0.$$

Set $\lambda_0 = c_2$. Then the Lax–Milgram lemma implies (i), (ii). Since $M_{k,\rho}$ is continuous it follows that the dual mapping $M^*_{k,\rho} : L^*_2 \to (\overset{0}{W}{}^{k,\rho})^*$ is also continuous; hence $\overline{M}_{k,\rho} = M^*_{k,\rho} \circ M_{k,\rho}$ is compact. Identifying $L_2 = L^*_2$, it is easy to see that $M^*_{k,\rho} f = f\rho^{-k}$; hence $\overline{M}_{k,\rho} f = \rho^{-2k} f$. Define $L := A + \lambda\overline{M}$, $\lambda > \lambda_0$. Then $A = L - \lambda\overline{M}$. Since L is an isomorphism by (ii) and $\overline{M}$ is compact, (iii) follows. Now the Riesz–Schauder theory applies to the operator $I + \mu L^{-1} \circ \overline{M}$. This implies (iv). Q.E.D.

4. The Neumann problem on unbounded domains. In order to study the Neumann problem by the variational method, we need suitable variants of the Poincaré lemma. On bounded domains we have (see Nečas [12], Wloka [18])

$$\| f \|^2_{W^k} \leqq \text{const} \left(\sum_{|\alpha|=k} \int |D^{\alpha} f|^2 dx + \sum_{|\alpha|<k} \left| \int D^{\alpha} f \, dx \right|^2 \right).$$

We first apply the compactness theorems of the last section to get a generalization for weighted spaces. Then we get it for convex domains and compute the constants explicitly. Furthermore, we prove generalizations of Friedrich's inequality. As an application we study the Neumann problem. Before we begin with this program, let us fix a notation. By Q_l, $l \in N_0$ we denote the set of all polynomials in $x \in R^n$ of degree less than or equal to l. For a multiindex $\alpha \in (N_0)^n$ we denote by q_a the monom $q_{\alpha}(x) := x^{\alpha}$.

THEOREM 4.1. (i) *Let* Ω, k, n, ρ *be such that* $M_{j,\rho} : W^{j,\rho}(\Omega) \to L_2(\Omega)$ *are compact for* $1 \leqq j \leqq k$. *Furthermore, let* $\rho^{-l} \in L_2(\Omega)$ *for some* $l \geqq 1$ *and let there exist a set* $I \subset \{ \alpha \in (N_0)^n \mid |\alpha| < k \}$ *such that* $W^{k,\rho}(\Omega) \cap Q_{k-1} = \text{span}\{ q_{\alpha} \mid \alpha \in I \}$. *Then there exists a constant* $c > 0$ *such that for all* $u \in W^{k,\rho}(\Omega)$

$$\| u \|^2_{k,\rho} \leqq c \left(\sum_{|\alpha|=k} \int_{\Omega} |D^{\alpha} u|^2 dx + \sum_{\alpha \in I} \left| \int_{\Omega} D^{\alpha} u \rho^{-k-l+|\alpha|} dx \right|^2 \right).$$

(ii) *The assertion remains true if we replace everywhere in* (i) $W^{k,\rho}$ *by* $W^{k,\rho}_c$.

Proof. (i) If the assertion is wrong there exists a sequence $u_m \in W^{k,\rho}$, $\|u_m\|_{k,\rho}=1$ such that

$$1 > m\left(\sum_{|\alpha|=k}\int |D^\alpha u_m|^2\,dx + \sum_{\alpha\in I}\left|\int D^\alpha u_m \rho^{-k-l+|\alpha|}\,dx\right|^2\right). \tag{1}$$

This implies that $\|D^\alpha u_m\|_2 \to 0$ for $m\to\infty$, $|\alpha|=k$. As in Theorem 3.17 we get functions g_α such that $\|D^\alpha u_m \rho^{-k+|\alpha|} - g_\alpha\|_2 \to 0$ for $m\to\infty$ and $|\alpha|<k$. We define $h_\alpha := g_\alpha \rho^{k-|\alpha|}$ for $|\alpha|<k$ and $h_\alpha = 0$ for $|\alpha|=k$. Then $D^\alpha h_0 = h_\alpha$ for $|\alpha|<k$ and from $D^\alpha h_0 = 0$ for $|\alpha|=k$ we get $h_0 \in Q_{k-1}$. Since $\rho^{-l}\in L_2$ (1) implies that

$$\int D^\alpha h_0 \rho^{-k-l+|\alpha|}\,dx = 0 \quad \text{for } \alpha\in I. \tag{2}$$

Since $h_0 = \lim u_m$ (in $W^{k,\rho}$) we see that on the one hand $\|h_0\|_{k,\rho}=1$ and on the other hand $h_0 = \sum_{\alpha\in I} a_\alpha q_\alpha$, $a_\alpha \in R$; hence (2) implies $h_0 = 0$. This is the desired contradiction.

(ii) Obvious. Q.E.D.

Now we are going to prove another variant of this type of inequality where we shall be able to compute the constants explicitly.

THEOREM 4.2. *Let Ω be a convex domain, ρ a convex function and $\rho \geq 1+|x|$, $\rho^{-1}|x| \in L_2$ and set $\rho_1 := \||x|\rho^{-1}\|_2^2$, $\rho_2 := \|\rho^{-1}\|_2^2$; $\rho_3 := \|\rho^{-k}\|_2^2$. Then for all $u \in W^{k,\rho}(\Omega)$*

$$\int_\Omega (u\rho^{-k})^2\,dx \leqq \frac{2^n(\rho_1+\rho_2)}{\rho_3}\int_\Omega \frac{|\nabla u|^2}{\rho^{2(k-1)}}\,dx + \frac{1}{\rho_3}\left(\int_\Omega \frac{u}{\rho^{2k}}\,dx\right)^2.$$

Proof. From $u(y)-u(x) = \int_0^1 \langle \nabla u(x+t(y-x))\,|\,(y-x)\rangle\,dt$ we get

$$u(y)^2 - 2u(y)u(x) + u(x)^2 \leqq \int_0^1 |\nabla u(x+t(y-x))|^2\,|y-x|^2\,dt.$$

Multiplying by $\rho^{-2k}(x)\rho^{-2k}(y)$ and integrating with respect to x, y yields

$$\begin{aligned} 2\int\rho^{-2k}\,dx\int (u\rho^{-k})^2\,dx - 2\left(\int u\rho^{-2k}\,dx\right)^2 \\ \leqq 2\int_0^1\int\int \frac{|\nabla u(x+t(y-x))|^2|x|^2}{\rho^{2k}(x)\rho^{2k}(y)}\,dx\,dy\,dt \\ + 2\int_0^1\int\int \frac{|\nabla u(x+t(y-x))|^2|y|^2}{\rho^{2k}(x)\rho^{2k}(y)}\,dx\,dy\,dt. \end{aligned} \tag{3}$$

We estimate the first summand in (3); the second is treated analogously. Set $\Phi(t,x,y) := x+t(y-x)$; then from the assumptions on ρ it follows that

$$\begin{aligned} &\int_0^{1/2}\int\int \frac{|\nabla u(\Phi(t,x,y))|^2|x|^2}{\rho^{2k}(x)\rho^{2k}(y)}\,dx\,dy\,dt \\ &\leqq \int_0^{1/2}\int \rho^{-2}(y)\int \frac{|\nabla u(\Phi(t,x,y))|^2}{(\rho(x)\rho(y))^{2(k-1)}}\,dx\,dy\,dt \\ &\leqq \int_0^{1/2}\int \rho^{-2}(y)\int \frac{|\nabla u(\Phi(t,x,y))|^2}{\rho(\Phi(t,x,y))^{2(k-1)}}\,dx\,dy\,dt =: I_1. \end{aligned}$$

The mapping $\Phi(t,\cdot,y):\Omega\to\Phi(t,\Omega,y)=:\Omega_{t,y}\subset\Omega$ is differentiable and bijective for all $t\in[0,\frac{1}{2}]$, $y\in\Omega$. The inverse is $\Phi(t,\cdot,y)^{-1}(x):=(z-ty)/(1-t)$. For the determinant of the Jacobian of this inverse we have $|J_{t,y}^{-1}|=(1-t)^{-n}\leqq 2^n$ for all $t\in[0,\frac{1}{2}]$. We transform the coordinates and obtain

$$I_1=\int_0^{1/2}\int_\Omega \rho^{-2}(y)\int_{\Omega_{t,y}}\frac{|\nabla u(z)|^2}{\rho(z)^{2(k-1)}}\left|J_{t,y}^{-1}\right|dz\,dy\,dt$$

$$\leqq 2^n\int_0^{1/2}\int_\Omega \rho^{-2}(y)\int\frac{|\nabla u(z)|^2}{\rho(z)^{2(k-1)}}\,dz\,dy\,dt$$

$$=2^{n-1}\rho_2\int\frac{|\nabla u(z)|^2}{\rho(z)^{2(k-1)}}\,dz.$$

Furthermore, we have

$$\int_{1/2}^1\int\int\frac{|\nabla u(\Phi(t,x,y))|^2|x|^2}{\rho^{2k}(x)\rho^{2k}(y)}\,dy\,dx\,dt$$

$$\leqq\int_{1/2}^1\int\frac{|x|^2}{\rho^2(x)}\int\frac{|\nabla u(\Phi(t,x,y))|^2}{\rho(\Phi(t,x,y))^{2(k-1)}}\,dy\,dx\,dt=:I_2.$$

Now we consider the transformation $\Phi(t,x,\cdot)$. For the determinant of the Jacobian of the inverse we have $|J_{t,x}^{-1}|=t^{-n}\leqq 2^n$ for all $t\in[\frac{1}{2},1]$ and $x\in\Omega$. A similar reasoning as above implies

$$I_2\leqq 2^{n-1}\rho_1\int\frac{|\nabla u(z)|^2}{\rho^{2(k-1)}(z)}\,dz.$$

The same estimates hold for the second summand in (3). Adding these estimates we obtain the conclusion. Q.E.D.

COROLLARY 4.3. *Let Ω, ρ be as above; then there exists a constant c_k such that for all $u\in W^{k,\rho}(\Omega)$*

$$\|u\|_{k,\rho}^2\leqq c_k\left(\sum_{|\alpha|=k}\|D^\alpha u\|_2^2+\sum_{|\alpha|<k}\left(\int_\Omega D^\alpha u\rho^{-2k+2|\alpha|}dx\right)^2\right).$$

Proof. Simple induction. Q.E.D.

We now turn to Friedrich's inequality. In this inequality appear some boundary integrals. So before we can proceed, we need a trace theorem for weighted spaces in order to ascertain that these integrals exist.

DEFINITION 4.4. The boundary $\partial\Omega$ of a set $\Omega\subset R^n$ is called uniformly locally finite Lipschitz continuous (ulf L) if there exist numbers L, α, $\beta>0$ and $K\in N$ as well as a covering $\{U_i\}_{i\in N}$ of $\partial\Omega$ such that

(i) At most K different U_i have a nonvoid intersection, i.e., $U_{j(1)},\cdots,U_{j(K+1)}\in\{U_i\}$ implies $\bigcap_{k=1}^{K+1}U_{j(k)}=\varnothing$.

(ii) For all U_i there exists an orthogonal transformation $A_i:R^n\to R^n$ ($y_1,\cdots,y_n$ are the new coordinates) and a uniformly Lipschitz continuous (with Lipschitz constant

L, independent of i) mapping $a_i: R^{n-1} \to R$ such that for $W_\alpha := \{(y_1, \cdots, y_{n-1}) \mid |y_j| \leqq \alpha\}$ we have

$$U_i \cap \partial\Omega = \{(y_1, \cdots, y_n) \mid (y_1, \cdots, y_{n-1}) \in W_\alpha,\ y_n = a_i(y_1, \cdots, y_{n-1})\},$$

$$U_i \cap \Omega = \{(y_1, \cdots, y_n) \mid (y_1, \cdots, y_{n-1}) \in W_\alpha,$$
$$a_i(y_1, \cdots, y_{n-1}) < y_n < a_i(y_1, \cdots, y_{n-1}) + \beta\},$$

$$U_i \cap C\bar{\Omega} = \{(y_1, \cdots, y_n) \mid (y_1, \cdots, y_{n-1}) \in W_\alpha,$$
$$a_i(y_1, \cdots, y_{n-1}) - \beta < y_n < a_i(y_1, \cdots, y_{n-1})\}.$$

LEMMA 4.5. *Let B, $B_i \in \text{Borel}(R^n)$, μ be a measure on* Borel(R^n). *If $B = \bigcup_{i=1}^{\infty} B_i$ and at most K different B_i have a nonvoid intersection, then $\sum_1^\infty \mu(B_i) \leqq K\mu(B)$.*

Proof. Let $P(m) := \{P_j\}_{j \in J}$ be a pavement of $\bigcup_1^m B_i$, i.e., the P_j are pairwise disjoint and

(i) for all P_j exists a $B_{i(j)}$, such that $P_j \subset B_{i(j)}$,
(ii) $B_i = \bigcup_{P \in B_i} P$,
(iii) $\bigcup_1^m B_i = \bigcup_{P \in P(m)} P$.

Then it follows that

$$\sum_1^m \mu(B_i) = \sum_1^m \sum_{P \subset B_i} \mu(P) \leqq K \sum_{P \in P(m)} \mu(P) = K \cdot \mu\left(\bigcup_{P \in P(m)} P \right) \leqq K \cdot \mu(B)$$

for all $m \in N$ and hence the assertion. Q.E.D.

THEOREM 4.6. *Let Ω have a ulfL-boundary and $\rho \geqq 1$ be a td-function. Then there exists a continuous, linear mapping $T: W^{k,\rho}(\Omega) \to L_{2,\rho^{-k}}(\partial\Omega)$ such that for all functions $u \in C^k(\bar{\Omega}) \cap W^{k,\rho}(\Omega)$ there holds $Tv = v \mid_{\partial\Omega}$.*

Proof. We work with the new coordinates $(y_1, \cdots, y_n)$ and set $y' = (y_1, \cdots, y_{n-1})$. Now fix $i \in N$ and $U_i^+ := \{y \mid y' \in W_\alpha,\ a_i(y') < y_n < a_i(y') + \beta\}$ and let $u \in C^k(\bar{\Omega}) \cap W^{k,\rho}(\Omega)$ then for $(y', z) \in U_i^+$

$$u(y', a_i(y')) = -\int_{a_i(y')}^{z} D_n u(y', y_n)\, dy_n + u(y', z)$$

and hence

$$u(y', a_i(y'))^2 \leqq 2\beta \int_{a_i(y')}^{a_i(y') + \beta} |D_n u(y', y_n)|^2 dy_n + 2u(y', z)^2.$$

We divide by $\rho^{2k}(y', a_i(y'))$ and since $\rho \geqq 1$ is a td-function

$$\frac{u(y', a_i(y'))^2}{\rho^{2k}(y', a_i(y'))} \leqq 2c_1\beta \int_{a_i(y')}^{a_i(y') + \beta} \frac{|D_n u(y', y_n)|^2}{\rho^{2(k-1)}(y', y_n)}\, dy_n + 2c_2 \frac{u(y', z)^2}{\rho^{2k}(y', z)}.$$

Integrating over y' and z, we conclude

$$\beta \int_{W_\alpha} \frac{u(y', a_i(y'))^2}{\rho^{2k}(y', a_i(y'))}\, dy' \leqq 2c_1\beta^2 \int\int_{U_i^+} \frac{|D_n u(y', y_n)|^2}{\rho^{2(k-1)}(y', y_n)}\, dy_n dy'$$
$$+ 2c_2 \int\int_{U_i^+} \frac{u(y', z)^2}{\rho^{2k}(y', z)}\, dz\, dy'.$$

The surface element is $ds=(1+\sum_{j=1}^{n-1}(D_j a_i)^2)^{1/2}dy'$ and since $|D_j a_i|\leqq L$ (see Definition 4.4), A_i orthogonal we obtain after transforming back to the old coordinates

$$\int_{U_i\cap\partial\Omega} u^2\rho^{-2k}ds\leqq c_3\|u\|^2_{W^{k,\rho}(U_i\cap\Omega)}$$

and hence (c_3 is independent of i)

$$\int_{\partial\Omega} u^2\rho^{-2k}ds\leqq c_3\sum_i\|u\|^2_{W^{k,\rho}(U_i\cap\Omega)}.$$

Lemma 4.5 implies the assertion. Q.E.D.

THEOREM 4.7. (i) *Let Ω have ulfL-boundary and $M_{j,\rho}: W^{j,\rho}(\Omega)\to L_2(\Omega)$ be compact for $1\leqq j\leqq k$. Furthermore, let $\rho^{-l}\in L_2$ for some $l\geqq 1$, $I\subset\{\alpha\in(N_0)^n\,|\,|\alpha|<k\}$ such that $Q_{k-1}\cap W^{k,\rho}(\Omega)=\operatorname{span}\{q_\alpha\,|\,\alpha\in I\}$ and there exists no polynomial which vanishes identically on $\partial\Omega$. Then*

$$\left(\sum_{|\alpha|=k}\int_\Omega|D^\alpha u|^2dx+\int_{\partial\Omega}u^2\rho^{-2k}ds\right)^{1/2}$$

is an equivalent Hilbert space norm on $W^{k,\rho}(\Omega)$.

(ii) *If we replace everywhere $W^{k,\rho}$ by $W_c^{k,\rho}$ the same statement holds.*

Proof. (i) From Theorem 4.1 we see that $(\sum_{|\alpha|=k}\|D^\alpha u\|_2^2)^{1/2}$ is equivalent to the quotient norm on $W^{k,\rho}(\Omega)/Q_{k-1}$ and from the assumptions on $\partial\Omega$ it follows that $\int_{\partial\Omega}u^2\rho^{-2k}ds$ is a norm on the finite dimensional space $Q_{k-1}\cap W^{k,\rho}(\Omega)$. This readily implies that $\|u\|^2:=\sum_{|\alpha|=k}\|D^\alpha u\|_2^2+\int_{\partial\Omega}u^2\rho^{-2k}ds$ is a Hilbert space norm on $W^{k,\rho}(\Omega)$—it remains only to prove the completeness of $W^{k,\rho}$ with respect to this norm. Since the imbedding $(W^{k,\rho}(\Omega),\|\cdot\|_{k,\rho})\to(W^{k,\rho}(\Omega),\|\cdot\|)$ is continuous by Theorem 4.6 and bijective, the assertion follows from the isomorphism theorem of Banach.

(ii) Obvious. Q.E.D.

Now we shall prove variants of Friedrich's inequality without using compactness theorems. Again we shall be able to compute the constants.

LEMMA 4.8. *Let Ω have ulfL-boundary and denote for any $x\in\partial\Omega$ by $v(x)=(v_1(x),\cdots,v_n(x))$ the normed outer normal. We assume that there exists $M>0$ such that $|\langle v(x),(x)\rangle|\leqq M$ for all $x\in\partial\Omega$. Now let $\rho\geqq|x|$, $\delta>0$, $k\in N$. Then either of the following two conditions*

$$\text{(i)}\qquad 1-\frac{2k\sum D_i\rho x_i}{n\rho}\geqq\delta>0\quad\text{for all } x\in\Omega,$$

$$\text{(ii)}\qquad 1-\frac{2k\sum D_i\rho x_i}{n\rho}\leqq-\delta<0\quad\text{for all } x\in\Omega$$

implies that for all $u\in W_c^{k,\rho}(\Omega)$

$$\int_\Omega\left(\frac{u}{\rho^k}\right)^2dx\leqq\frac{4}{(n\delta)^2}\sum_1^n\int_\Omega\left|\frac{D_iu}{\rho^{k-1}}\right|^2dx+\frac{2M}{n\delta}\int_{\partial\Omega}\left(\frac{u}{\rho^k}\right)^2ds.$$

Proof. Let $u\in C_0^\infty(R^n)\,|_\Omega$. Then by Gauss' integral theorem

$$\int_\Omega\left(\frac{u}{\rho^k}\right)^2dx=-\frac{2}{n}\sum_1^n\int_\Omega\left[\frac{D_iu}{\rho^k}-\frac{kuD_i\rho}{\rho^{k+1}}\right]\frac{u}{\rho^k}x_idx+\frac{1}{n}\sum_1^n\int_{\partial\Omega}v_i(x)\left(\frac{u}{\rho^k}\right)^2x_ids$$

and hence

(4)

$$\int_\Omega \left(\frac{u}{\rho^k}\right)^2 \left[1 - \frac{2k\sum D_i \rho x_i}{n\rho}\right] dx = -\frac{2}{n}\sum \int_\Omega \frac{D_i u}{\rho^{k-1}} \cdot \frac{u}{\rho^k} \cdot \frac{x_i}{\rho}\, dx + \int_{\partial\Omega} \frac{\langle v(x), x\rangle}{n} \left(\frac{u}{\rho^k}\right)^2 ds.$$

(i) In this case (4) implies

$$\int_\Omega \left(\frac{u}{\rho^k}\right)^2 dx \leqq \frac{2}{n\delta}\sum \int_\Omega \left|\frac{D_i u}{\rho^{k-1}} \cdot \frac{u}{\rho^k}\right| dx + \frac{M}{n\delta}\int_{\partial\Omega}\left(\frac{u}{\rho^k}\right)^2 ds$$

$$\leqq \frac{2}{(n\delta)^2}\sum \int_\Omega \left(\frac{D_i u}{\rho^{k-1}}\right)^2 dx + \frac{1}{2}\int_\Omega \left(\frac{u}{\rho^k}\right)^2 dx + \frac{M}{n\delta}\int_{\partial\Omega}\left(\frac{u}{\rho^k}\right)^2 ds$$

and hence the assertion.

(ii) We estimate in (4) on the left side from above by $-\delta \int u^2 \rho^{-2k}\, dx$, divide by $-\delta$ and proceed as in (i). Q.E.D.

Remark 4.9. If Lemma 4.8, case (i) for all $x \in \partial\Omega$ $\langle v(x), x\rangle \leqq 0$—as is the case e.g. for $S(0, r_0, \Sigma)$—we can omit the boundary integrals in the assertion of Lemma 4.8. To see this, consider again the identity (4).

THEOREM 4.10. *Let Ω fulfill the conditions of Lemma* 4.8. *Furthermore, let Ω, ρ, j, n for all $j \leqq k$ fulfill one of the conditions of Theorem* 2.3. *Then for all $u \in W_c^{k,\rho}(\Omega)$*

$$\sum_{|\alpha|=k} \|D^\alpha u\|_2^2 \leqq \|u\|_{k,\rho}^2 \leqq \sum_{|\alpha|=k_0} \|D^\alpha u\|_2^2 + \sum_{|\alpha|<k} c_\alpha \int_{\partial\Omega}\left(\frac{D^\alpha u}{\rho^{k-|\alpha|}}\right)^2 ds.$$

Proof. By induction from Lemma 4.8 and the proof of Theorem 2.3. Q.E.D.

As an application we consider the following Neumann problem for an elliptic operator of order $2k$: Find $u \in W^{k,\rho}(\Omega)$ such that for all $v \in W^{k,\rho}(\Omega)$

$$(5) \qquad a(u,v) := \sum_{|\alpha|=|\beta|=k} \int_\Omega a_{\alpha\beta} D^\beta u(x) D^\alpha v(x)\, dx = \int_\Omega f(x) v(x)\, dx.$$

THEOREM 4.11. *Let $a_{\alpha\beta} \in L_\infty$, $(a_{\alpha\beta})_{|\alpha|=|\beta|=k}$ uniformly positive definite and $f\rho^k \in L_2(\Omega)$. Furthermore, let ρ be convex and $\rho^{-1}\cdot(1+|x|) \in L_2(R^n)$. We assume that there exists a convex set $\tilde{\Omega}$ and a (l,κ)-diffeomorphism $T: \tilde{\Omega} \to \Omega$ with $1 \leqq l+\kappa$ such that the matrix*

$$B := \left(B_{\alpha\beta}(T)\right)_{|\alpha|,|\beta|<k},\ B_{\alpha\beta}(T) := \int_{\tilde{\Omega}} D^\alpha\left(q_\beta \circ T\right)(y)\rho^{-2k+|\alpha|}(y)\, dy$$

is invertible. Then problem (5) *admits a solution u iff for all $q \in Q_{k-1}$ $\int_\Omega q \cdot f\, dx = 0$. This solution is unique in $W^{k,\rho}(\Omega)/Q_{k-1}$ and*

$$\|u\|_{W^{k,\rho}/Q_{k-1}} \leqq c\|f\rho^k\|_{L_2},$$

where c can be computed explicitly.

Proof. $\rho^{-1}(1+|x|) \in L_2$ implies $Q_{k-1} \subset W^{k,\rho}$ is well defined. Denote by $Q_{k-1}^\perp$ the orthogonal complement of Q_{k-1} then $Q_{k-1}^\perp \cong W^{k,\rho}/Q_{k-1}$ and the norms can be identified. Hence we may denote the norm on the quotient space by $\|\cdot\|_{k,\rho,\perp}$. Applying Lemma 3.6, Theorem 4.2 and Lemma 3.2, we obtain for all $v \in W^{k,\rho}(\Omega)$

$$\|v\|_{k,\rho}^2 \leqq c_1\left(\sum_{|\alpha|=k} \|D^\alpha v\|_2^2 + \sum_{|\alpha|<k}\left|\int_{\tilde{\Omega}} D^\alpha(v\circ T)(y)\rho^{-2k+|\alpha|}(y)\, dy\right|^2\right).$$

If $q := \Sigma_{|\beta|<k} d_\beta q_\beta \in Q_{k-1}$, $d_\beta \in R$, then

$$\|v\|^2_{k,\rho,\perp} \leqq \|v+q\|^2_{k,\rho}$$

$$\leqq c_1 \left(\sum_{|\alpha|=k} \int_\Omega |D^\alpha v|^2 dx + \sum_{|\alpha|<k} \left| c_\alpha(v) + \sum_{|\beta|<k} d_\beta B_{\alpha\beta}(T) \right|^2 \right)$$

where $c_\alpha := \int_{\tilde{\Omega}} D^\alpha (v \circ T)(y) \rho^{-2k+|\alpha|}(y)\, dy$. Since B is invertible there exists a vector d such that $-c(v) := Bd$ and hence

$$\|v\|^2_{k,\rho,\perp} \leqq c_1 \sum_{|\alpha|=k} \int_\Omega |D^\alpha v|^2 dx.$$

Now let u be a solution of (5). Then for all $q \in Q_{k-1}$ and all $v \in W^{k,\rho}(\Omega)$

$$a(u,v) = a(u, v+q) = \int fv\, dx + \int fq\, dx$$

and $\int fq\, dx = 0$. If on the other hand $\int fq\, dx$ for all $q \in Q_{k-1}$ then $\int fv\, dx$ is a continuous, linear functional on $Q^\perp_{k-1}$. Since there appear only derivations of order k $a(u,v)$ is a continuous, bilinear form on $Q^\perp_{k-1}$. Furthermore, $a(u,u) \geqq c_2 \Sigma_{|\alpha|=k} \|D^\alpha u\|^2_2 \geqq c_2 c_1^{-1} \|u\|^2_{k,\rho,\perp}$ and the Lax–Milgram lemma concludes the proof. Q.E.D.

Remark 4.12. (i) Analogously, we might apply the other inequalities to get existence theorems for problem (5). In some cases the decay condition $u \in W^{k,\rho}_c$ suffices to ensure the existence of a unique solution, e.g. in the situation of Theorems 4.10, 4.1 if $Q_{k-1} \cap W^{k,\rho}(\Omega) = \{0\}$.

(ii) The coercivity estimates can be used to prove the convergence of approximate solutions computed on bounded domains. Also they can be applied to get stochastic representations of solutions to the second-order Dirichlet problem (see Janßen [7]).

REFERENCES

[1] R. A. ADAMS, *Sobolev spaces*, in Pure and Applied Mathematics 65, Academic Press, New York, 1975.

[2] V. BENCI AND D. FORTUNATO, *Some compact embedding theorems for weighted Sobolev spaces*, Boll. Un. Math. Ital. Sect. B, Ser. 5, 13 (1976), pp. 832–843.

[3] O. V. BESOV ET AL., *Imbedding theory for differentiable functions*, AMS Trans. Ser. 2, 105 (1976), pp. 57–94.

[4] M. Cantor, *Elliptic operators and the decomposition of tensor fields*, Bull. Amer. Math. Soc., 5 (1981), pp. 235–262.

[5] R. JANßEN, *Some variants of the Poincaré lemma*, Appl. Anal., 14 (1983), pp. 303–315.

[6] _____, *The Dirichlet problem for second order elliptic operators on unbounded domains*, Appl. Anal., 19 (1985), pp. 201–216.

[7] _____, *Stochastic representation of solutions to the Dirichlet problem on unbounded domains and the decay at infinity*, submitted to Stoch. Proc. Appl.

[8] L. D. KUDRJAVCEV, *The solution of the first boundary value problem for self-adjoint elliptic equations in the case of an unbounded region*, Math. USSR Izvestija, 1 (1967), pp. 1131–1151.

[9] _____, *A variational method for unbounded regions*, Sov. Math. Dokl., 5 (1964), pp. 887–890.

[10] J. MAULEN, *Lösungen der Poissongleichung und harmonische Vektorfelder in unbeschränkten Gebieten*, Math. Meth. Appl. Sci., 5 (1983), pp. 233–255.

[11] W. MAZJA, *Einbettungssätze für Sobolevsche Räume*, vol. I, Teubner, Leipzig, 1979.

[12] J. NEČAS, *Les méthodes directes en théorie des équations élliptiques*, Masson et Cie, Paris, 1967.

[13] J. T. ODEN AND J. N. REDDY, *The Mathematical Theory of Finite Element Methods*, Springer, Berlin, 1983.

[14] M. OTELBAEV, *Imbedding theorems for weighted Sobolev spaces and their applications in the study of the spectrum of the Schrödinger operator*, Trudy Mat. Inst. Steklov. 150 (1979), pp. 265–305 = Proc. Steklov Inst. Math., 4 (1981), pp. 281–318.

[15] OWEN, *Boundary value problems for the Laplacian in an exterior domain*, Comm. PDE, 6/7 (1981), pp. 783–798.

[16] V. VOGELSANG, *Elliptische Differentialgleichungen mit variablen Koeffizienten in Gebieten mit unbeschränktem Rand*, Manuscripta Math., 14 (1975), pp. 379–401.

[17] A. VOIGT AND J. WLOKA, *Hilberträume und elliptische Differentialoperatoren*, Bibliographisches Institut, Mannhiem, 1975.

[18] J. WLOKA, *Partielle Differentialgleichungen*, Teubner, Leipzig, 1982.

[19] K. YOSHIDA, *Functional Analysis*, Springer-Verlag, New York, 1966.

SIAM J. MATH. ANAL.
Vol. 17, No. 6, November 1986

008

EXISTENCE AND UNIQUENESS OF SOLUTIONS TO THE COMPRESSIBLE REYNOLDS LUBRICATION EQUATION*

MICHEL CHIPOT† AND MITCHELL LUSKIN‡

Abstract. We prove the existence of a solution to the compressible Reynolds lubrication equation and we show that our solution is unique in the class of nonnegative solutions (under some additional hypotheses, we prove that our solution is unique among all weak solutions). We also prove the strong result that the mapping from the boundary data to the solution is monotonic.

Key words. compressible Reynolds lubrication equation, nonlinear elliptic boundary value problem

AMS(MOS) subject classifications. Primary 76N99, 35J65

1. Introduction. The compressible Reynolds lubrication equation [7, p. 63] is the nonlinear elliptic partial differential equation

$$\begin{aligned} &6\mu \mathbf{V}\cdot\nabla(Ph)=\nabla\cdot(h^3P\nabla P), \qquad x=(x_1,x_2)\in\Omega,\\ &P=P_a, \qquad x\in\partial\Omega, \end{aligned} \tag{1.1}$$

which gives the pressure, $P=P(x)$, that develops in a layer of air of thickness, $h=h(x)$, which is confined between two solid bodies when the average of the velocities of the upper and lower bodies is $\mathbf{V}=(V_1,V_2)$. The air is assumed to be isothermal and to be an ideal gas (the density is taken to be proportional to the pressure) [7]. Here, $\mu>0$ is the dynamic viscosity of the fluid, $\Omega\subseteq\mathbb{R}^2$ is the region (with smooth boundary, $\partial\Omega$) where the upper and lower bodies are in proximity, and $P_a>0$ is the ambient pressure.

When the thickness of a gaseous fluid layer is of the order of the molecular mean free path of the gas, then the compressible Reynolds equation becomes a poor model for the pressure in the fluid layer. In many applications in the modeling of the mechanical systems of magnetic recording the following modified Reynolds equation [3] has been found to be a good model equation for predicting the pressure in the fluid layer

$$\begin{aligned} &6\mu \mathbf{V}\cdot\nabla(Ph)=\nabla\cdot(h^3P\nabla P)+\nabla\cdot\left(6\lambda_a h^2P_a\nabla P\right), \qquad x\in\Omega,\\ &P=P_a, \qquad x\in\partial\Omega. \end{aligned} \tag{1.2}$$

Here $\lambda_a\geqq 0$ is the mean free path of the gas at ambient pressure. Note that $\lambda_a=0$ gives the compressible Reynolds equation (1.1) and all the analysis and results which we give for the modified Reynolds equation, (1.2), are valid for the modified Reynolds equation, (1.1), unless we state otherwise.

* Received by the editors January 28, 1985, and in revised form August 16, 1985.

† Institute for Mathematics and Its Applications, University of Minnesota, Minneapolis, Minnesota 55455. On leave from Universite de Nancy I, Département de Mathématiques, BP 239-54506, Vandoeuvre Cedex, France.

‡ School of Mathematics, University of Minnesota, Minneapolis, Minnesota 55455. The research of this author was supported by the National Science Foundation under grants DMS 830-1575 and DMS 835-1080.

We assume only that $h=h(x)$ is a Lipschitz continuous function such that for positive constants $h_{\min}, h_{\max}$, and H we have the bounds

$$\begin{aligned} &0<h_{\min}\leqq h(x)\leqq h_{\max}, \qquad x\in\Omega, \\ &|\nabla h(x)|\leqq H, \qquad \text{a.e. } x\in\Omega. \end{aligned} \tag{1.3}$$

We prove the existence of a nonnegative weak solution to the Reynolds equation (1.2), and we prove that this solution is unique in the class of nonnegative functions. We actually prove the stronger result that the mapping S: $P_a\to P(x)$ from the boundary data to the solution satisfies the monotonicity result

$$P_a\geqq Q_q \text{ implies } P(x)\geqq Q(x) \quad \text{for all } x\in\Omega$$

where $Q=S(Q_a)$. Moreover, if

$$\mathbf{V}\cdot\nabla h\leqq 0,$$

we prove that our solution to the Reynolds equation (1.1) is unique among all weak solutions.

We note that the implicit function theorem has been used in [5] to obtain the existence of solutions to (1.1) for small values of the velocity, $V=(V_1^2+V_2^2)^{1/2}$. However, our techniques give the existence and uniqueness of solutions to (1.1) for all values of $\mathbf{V}$.

In the mechanical systems of magnetic recording, the pressure developed in the fluid layer is coupled to the deformation of the confining solid bodies (such as a disk or a tape) [8]. In a forthcoming paper, we will give an analysis of the system of coupled partial differential equations for the pressure and the deformation. In this case, h will depend on the deformation and the present results will be useful there.

2. Existence of solutions. In what follows, the $L^2(\Omega)$ inner product for real-valued functions ψ, $\zeta\in L^2(\Omega)$ is denoted by

$$(\psi,\zeta)=\int_\Omega \psi(x)\zeta(x)\,dx$$

with corresponding norm

$$\|\psi\|^2=(\psi,\psi).$$

If ψ, ζ: $\Omega\to\mathbb{R}^2$, then the $L^2(\Omega)$ inner product for ψ, $\zeta\in L^2(\Omega)$ is similarly denoted by

$$(\psi,\zeta)=\int_\Omega \psi(x)\cdot\zeta(x)\,dx$$

with corresponding norm

$$\|\psi\|^2=(\psi,\psi),$$

where $\psi(x)\cdot\zeta(x)$ denotes the usual Euclidean inner product in $\mathbb{R}^2$. Also, we define the Sobolev spaces

$$\begin{aligned} H^1(\Omega)&=\{\psi\in L^2(\Omega)\mid\nabla\psi\in L^2(\Omega)\}, \\ H_0^1(\Omega)&=\{\psi\in H^1(\Omega)\mid\psi\equiv 0 \text{ on } \partial\Omega\}, \end{aligned}$$

with norm

$$\|\psi\|^2_{H^1(\Omega)} = \|\psi\|^2 + \|\nabla\psi\|^2,$$

where equality on the boundary is understood in the trace sense [1].

We first prove the existence of a solution to (1.2). Set $\Lambda = 6\mu V$ and $\lambda = 6\lambda_a P_a$ and introduce the dependent variable

$$u = \frac{1}{2}P^2 + \frac{\lambda}{h}P.$$

If P is a nonnegative solution to (1.2), then we see that u is a nonnegative solution to

$$\begin{aligned} &\nabla\cdot(h^3\nabla u) = \nabla\cdot(\beta(x,u)\Lambda - \lambda\beta(x,u)\nabla h(x)), \qquad x\in\Omega, \\ &u(x) = \frac{1}{2}P_a^2 + \frac{\lambda}{h(x)}P_a \equiv u_a(x), \qquad x\in\partial\Omega, \end{aligned} \tag{2.1}$$

where

$$\beta(x,u) = \begin{cases} -\lambda + \sqrt{\lambda^2 + 2h^2 u}, & u \geqq 0, \\ 0, & u \leqq 0. \end{cases}$$

Conversely, if u is a nonnegative of (2.1), then

$$P(x) = -\frac{\lambda}{h(x)} + \sqrt{\lambda^2/h(x)^2 + 2u(x)}$$

is a nonnegative solution to (1.2). Thus, we shall show that (2.1) has a nonnegative solution.

We set

$$\alpha(x,u) = \beta(x,u)\Lambda - \lambda\beta(x,u)\nabla h(x).$$

We define a weak solution, u, to the problem:

$$\begin{aligned} &\nabla\cdot(h^3(x)\nabla u) = \nabla\cdot\alpha(x,u), \qquad x\in\Omega, \\ &u = \varphi, \qquad x\in\partial\Omega, \end{aligned} \tag{2.2}$$

for $\varphi\in H^1(\Omega)$ to be u such that

$$\begin{aligned} &u - \varphi \in H_0^1(\Omega), \\ &\int_\Omega h^3(x)\nabla u\cdot\nabla\xi\,dx = \int_\Omega \alpha(x,u)\cdot\nabla\xi\,dx, \qquad \xi\in H_0^1(\Omega). \end{aligned} \tag{2.3}$$

Using the inequality $\sqrt{A} + \sqrt{B} \geqq \sqrt{A+B}$, it is easy to check that

$$\begin{aligned} &|\beta(x,v)|^2 \leqq 2h^2|v|, \qquad x\in\Omega, \quad v\in\mathbb{R}, \\ &|\beta(x,v) - \beta(x,w)|^2 \leqq 2h^2|v-w|, \qquad x\in\Omega, \quad v,w\in\mathbb{R}. \end{aligned}$$

Hence for some constants c_1, c_2 we have

$$\begin{aligned} &|\alpha(x,v)|^2 \leqq c_1|v|, \qquad x\in\Omega, \quad v\in\mathbb{R}, \\ &|\alpha(x,v) - \alpha(x,w)|^2 \leqq c_2|v-w|, \qquad x\in\Omega, \quad v,w\in\mathbb{R}. \end{aligned} \tag{2.4}$$

We shall prove the following theorem.

THEOREM 1. *There exists a weak solution to* (2.2).

In §3, we shall prove that a (weak) solution u, to (2.1) is nonnegative almost everywhere. Thus, we can define

$$P(x)=\frac{-\lambda}{h(x)}+\sqrt{\lambda^2/h(x)^2+2u(x)} \tag{2.5}$$

to be a weak solution to (1.2). Hence, we have the following theorem.

THEOREM 2. *There exists a nonnegative, weak solution, P, to* (1.2).

We now turn to the proof of Theorem 1. We shall use the Schauder fixed point theorem. We denote by T: $L^2(\Omega)\to L^2(\Omega)$ the map $u=T(v)$ where $u\in H^1(\Omega)$ is the solution to $u=\varphi$ on $\partial\Omega$,

$$\int_\Omega h^3(x)\nabla u\cdot\nabla\xi\,dx=\int_\Omega \alpha(x,v)\cdot\nabla\xi\,dx,\qquad \xi\in H_0^1(\Omega). \tag{2.6}$$

The Schauder fixed point theorem states that if T is continuous and if there exists a closed, convex set B such that $T(B)\subset B$ and $\overline{T(B)}$ is compact, then there exists a fixed point, $u\in B$, of T, i.e., $T(u)=u$. We note that a fixed point of T is a weak solution of (2.2).

Set

$$B_R=\{v\in L^2(\Omega)\,|\,\|v\|\leqq R\}.$$

We shall show that there exists a positive constant R_1 such that

$$T(B_R)\subseteq B_R \quad \text{if } R\geqq R_1.$$

Further, we shall show that there exists a positive constant, $c_3=c_3(R)$, such that

$$\|T(v)\|_{H^1(\Omega)}\leqq c_3,\qquad v\in B_R. \tag{2.7}$$

Thus, it follows from Rellich's theorem [1] that $\overline{T(B_R)}$ is compact.

The conditions (2.6) are equivalent to the conditions

$$\begin{aligned} &u-\varphi\in H_0^1(\Omega),\\ &\int_\Omega h^3(x)\nabla(u-\varphi)\cdot\nabla\xi\,dx=-\int_\Omega h^3(x)\nabla\varphi\cdot\nabla\xi\,dx\\ &\qquad\qquad +\int_\Omega \alpha(x,v)\cdot\nabla\xi\,dx,\qquad \xi\in H_0^1(\Omega). \end{aligned} \tag{2.8}$$

It follows from standard elliptic theory [2] that (2.8) has a unique solution, $u\in H^1(\Omega)$. (Note that $\alpha(x,v)\in L^2(\Omega)$ thanks to (2.4).) We can set $\xi=u-\varphi$ in (2.8) and use the Cauchy-Schwarz inequality to obtain the bound (see (1.3))

$$\|\nabla(u-\varphi)\|\leqq c\big(\|\nabla\varphi\|+\|\alpha(x,v)\|\big). \tag{2.9}$$

Here and in what follows, c will denote a positive constant which can change from equation to equation. Now it follows from (2.4) that

$$\|\alpha(x,v)\|\leqq c_1^{1/2}\|v\|_{L^1(\Omega)}\leqq c_1^{1/2}|\Omega|^{1/2}\|v\|^{1/2}, \tag{2.10}$$

where $|\Omega|$ is the measure of Ω. Hence, we obtain from (2.9) and the triangle inequality

$$\|\nabla u\| \leqq \|\nabla \varphi\| + \|\nabla(u-\varphi)\| \leqq c\left(\|\nabla \varphi\| + \|v\|^{1/2}\right). \tag{2.11}$$

Also,

$$\|\psi\| \leqq \nu^{-1/2}\|\nabla \psi\|, \qquad \psi \in H_0^1(\Omega), \tag{2.12}$$

where $\nu > 0$ is the smallest eigenvalue of the problem

$$\begin{aligned} -\Delta\psi &= \nu\psi, && x \in \Omega, \\ \psi &= 0, && x \in \partial\Omega. \end{aligned}$$

Now

$$\|u\| \leqq \|\varphi\| + \|u-\varphi\| \leqq \|\varphi\| + \nu^{1/2}\|\nabla(u-\varphi)\|. \tag{2.13}$$

Thus, it follows from (2.9), (2.10), and (2.13) that there exists positive constants c_4 and c_5 such that

$$\|u\| \leqq c_4\|\varphi\|_{H^1(\Omega)} + c_5\|v\|^{1/2}. \tag{2.14}$$

We see from (2.14) that $\|v\| \leqq R$ implies that $\|u\| \leqq R$ if $R \geqq R_1$ where

$$R_1 = \left(\frac{c_5 + \sqrt{c_5^2 + 4c_4\|\varphi\|_{H^1(\Omega)}^2}}{2} \right).$$

Thus, $T(B_R) \subseteq B_R$ if $R \geqq R_1$. Further, it follows from (2.11) that there exists $c_3 = c_3(R)$ such that

$$\|T(v)\|_{H^1(\Omega)} \leqq c_3(R), \qquad v \in B_R.$$

All of the hypotheses of Schauder's theorem have now been satisfied except for the continuity of T. It follows from (2.6) that for $v, w \in L^2(\Omega)$,

$$T(v) - T(w) \in H_0^1(\Omega),$$

$$\int_\Omega h^3(x)\nabla(T(v)-T(w))\cdot\nabla\xi\,dx = \int_\Omega (\boldsymbol{\alpha}(x,v)-\boldsymbol{\alpha}(x,w))\cdot\nabla\xi\,dx, \qquad \xi \in H_0^1(\Omega). \tag{2.15}$$

Thus, we can set $\xi = T(v) - T(w)$ above to obtain

$$\|\nabla(T(v)-T(w))\| \leqq c\|\boldsymbol{\alpha}(x,v)-\boldsymbol{\alpha}(x,w)\|. \tag{2.16}$$

Now by (2.5)

$$\|\boldsymbol{\alpha}(x,v)-\boldsymbol{\alpha}(x,w)\| \leqq c_2^{1/2}\|v-w\|_{L^1(\Omega)} \leqq c_2^{1/2}|\Omega|^{1/2}\|v-w\|^{1/2}.$$

Hence, it follows from (2.12), (2.16), and (2.17) that

$$\|T(v)-T(w)\| \leqq \|v-w\|^{1/2},$$

i.e., T is Hölder continuous with exponent $1/2$. This completes the proof that T has a fixed point.

Remark. One could weaken the assumption (2.4) in various directions and still get a solution of (2.3).

3. Uniqueness of solutions. In this section, we prove a uniqueness and monotonicity result for weak solutions to the problem (2.2). More precisely, we have:

THEOREM 3. *The weak solution to* (2.2) *is unique. Further, suppose that* u_1 *is a weak solution to* (2.2) *corresponding to boundary data* φ_1 *and* u_2 *is a weak solution to* (2.2) *corresponding to boundary data* φ_2. *If* $\varphi_1 \geqq \varphi_2$ *a.e. on* $\partial\Omega$, *then* $u_1 \geqq u_2$ *a.e. in* Ω.

Proof. The uniqueness of weak solutions clearly follows from the monotonicity result. We will use here an argument due to Carillo and Chipot. See [4] for a variant.

We assume that $\varphi_1 \geqq \varphi_2$ a.e. on $\partial\Omega$. First, we prove that for all $\zeta \in C^\infty(\bar{\Omega})$ and $\zeta > 0$ we have

$$\int_{[u_1 - u_2 > 0]} h^3(x)\nabla(u_2 - u_1)\cdot\nabla\zeta - (\alpha(x,u_2) - \alpha(x,u_1))\cdot\nabla\zeta\, dx \leqq 0 \tag{3.1}$$

where

$$[u_2 - u_1 > 0] = \{x \in \Omega \mid u_2(x) - u_1(x) > 0\}.$$

To do that, we consider for $\varepsilon > 0$

$$\xi = \min\left(\frac{(u_2 - u_1)^+}{\varepsilon}, \zeta\right) \tag{3.2}$$

where

$$\psi^+(x) = \max(\psi(x), 0).$$

Note that $\xi \in H_0^1(\Omega)$ since for $x \in \partial\Omega$,

$$\xi(x) = \min\left(\frac{(\varphi_2 - \varphi_1)^+}{\varepsilon}, \zeta\right) = 0.$$

It follows from subtracting (2.3) with $u = u_2$ from (2.3) with $u = u_1$ that

$$\int_\Omega h^3(x)\nabla(u_2 - u_1)\nabla\xi - (\alpha(x,u_2) - \alpha(x,u_1))\cdot\nabla\xi\, dx = 0, \tag{3.3}$$

which for ξ given by (3.2) is equivalent to

$$\begin{aligned}\int_{[u_2 - u_1 > \varepsilon\zeta]} & h^3(x)\nabla(u_2 - u_1)\nabla\zeta - (\alpha(x,u_2) - \alpha(x,u_1))\cdot\nabla\zeta\, dx \\ & + \frac{1}{\varepsilon}\int_{[0 < u_2 - u_1 \leqq \varepsilon\zeta]} h^3(x)|\nabla(u_2 - u_1)|^2 dx \\ & - \frac{1}{\varepsilon}\int_{[0 < u_2 - u_1 \leqq \varepsilon\zeta]} (\alpha(x,u_2) - \alpha(x,u_1))\cdot\nabla(u_2 - u_1)\, dx = 0\end{aligned} \tag{3.4}$$

where

$$[u_2 - u_1 > \varepsilon\zeta] = \{x \in \Omega \mid u_2(x) - u_1(x) > \varepsilon\zeta\}$$

and the other sets $[\cdot]$ are defined likewise. We estimate the last integral above by

$$
\begin{aligned}
I &\equiv \int_{[0<u_2-u_1\leqq\varepsilon\zeta]} (\boldsymbol{\alpha}(x,u_2)-\boldsymbol{\alpha}(x,u_1))\cdot\nabla(u_2-u_1)\,dx \\
&\leqq \left(\int_{[0<u_2-u_1\leqq\varepsilon\zeta]} h(x)^{-3}|\boldsymbol{\alpha}(x,u_2)-\boldsymbol{\alpha}(x,u_1)|^2\,dx\right)^{1/2} \\
&\quad\times\left(\int_{[0<u_2-u_1\leqq\varepsilon\zeta]} h^3(x)|\nabla(u_2-u_1)|^2\,dx\right)^{1/2} \qquad (3.5)\\
&\leqq \frac{1}{4}\int_{[0<u_2-u_1\leqq\varepsilon\zeta]} h(x)^{-3}|\boldsymbol{\alpha}(x,u_2)-\boldsymbol{\alpha}(x,u_1)|^2\,dx \\
&\quad+\int_{[0<u_2-u_1\leqq\varepsilon\zeta]} h^3(x)|\nabla(u_2-u_1)|^2\,dx.
\end{aligned}
$$

Using the estimate (3.5) in (3.4) we obtain from (1.3) and (2.4)

$$
\begin{aligned}
&\int_{[u_2-u_1>\varepsilon\zeta]} h^3(x)\nabla(u_2-u_1)\nabla\zeta-(\boldsymbol{\alpha}(x,u_2)-\boldsymbol{\alpha}(x,u_1))\cdot\nabla\zeta\,dx \qquad (3.6)\\
&\qquad\leqq \frac{1}{4\varepsilon}\int_{[0<u_2-u_1\leqq\varepsilon\zeta]} h(x)^{-3}|\boldsymbol{\alpha}(x,u_2)-\boldsymbol{\alpha}(x,u_1)|^2\,dx \\
&\qquad\leqq \frac{c_2M}{4h_{\min}^3}\int_{[0<u_2-u_1\leqq\varepsilon\zeta]} dx
\end{aligned}
$$

where $M=\max\zeta$. Now the measure of the set $[0<u_2-u_1\leqq\varepsilon\zeta]$ goes to zero as $\varepsilon\to 0$. Thus, the estimate (3.1) follows from (3.6).

Now set $\mathbf{n}=(n_1,n_2)\equiv(-\Lambda_2,\Lambda_1)$ and $s>0$. We then set

$$
\zeta(x_1,x_2)=W-\exp(s(n_1x_1+n_2x_2)) \qquad (3.7)
$$

where W is a constant chosen large enough so that $\zeta>0$. If we set ζ from (3.7) in (3.1) we obtain (see the definition of $\boldsymbol{\alpha}$)

$$
\int_{[u_2-u_1>0]} h^3(x)\nabla(u_2-u_1)\nabla\zeta+\lambda(\beta(x,u_2)-\beta(x,u_1))\nabla h\cdot\nabla\zeta\,dx\leqq 0 \qquad (3.8)
$$

since $\boldsymbol{\Lambda}\cdot\nabla\zeta=0$ for all $x\in\Omega$. Now it follows from integration by parts that

$$
\begin{aligned}
\int_{[u_2-u_1>0]} h^3(x)\nabla(u_2-u_1)\nabla\zeta\,dx &= \int_\Omega h^3(x)\nabla(u_2-u_1)^+\nabla\zeta\,dx \qquad (3.9)\\
&= -\int_\Omega (u_2-u_1)^+\nabla\cdot(h^3(x)\nabla\zeta)\,dx \\
&= -\int_{[u_2-u_1>0]} (u_2-u_1)\nabla\cdot(h^3(x)\nabla\zeta)\,dx.
\end{aligned}
$$

So, from (3.8) we obtain

$$
\int_{[u_2-u_1>0]} (u_2-u_1)\left[-\nabla\cdot(h^3(x)\nabla\zeta)+g(x)\nabla h\cdot\nabla\zeta\right]dx\leqq 0
$$

where

$$g(x)=\begin{cases}\dfrac{\lambda(\beta(x,u_2(x)))-\beta(x,u_1(x))}{(u_2(x)-u_1(x))} & \text{if } u_2(x)\neq u_1(x),\\ 0 & \text{if } u_2(x)=u_1(x).\end{cases}$$

If $\lambda=0$, then $g=0$. Further, from the definition of β, for $\lambda\neq 0$

$$|\beta(x,v)-\beta(x,w)|\leqq\frac{h_{\max}^2}{\lambda}|v-w|,\quad x\in\Omega,\quad v,w\in\mathbb{R}.$$

Thus, $g\in L^\infty(\Omega)$. Also, we have that

$$\begin{aligned}-\nabla\cdot\big(h^3(x)\nabla\zeta\big)+g(x)\nabla h\cdot\nabla\zeta&\\ =-h^3(x)\Delta\zeta-\nabla h^3\cdot\nabla\zeta+g(x)\nabla h\cdot\nabla\zeta&\\ =\exp\big(s(n_1x_1+n_2x_2)\big)\big[h^3(x)s^2n^2+(\nabla h^3\cdot\mathbf{n})s-g(x)(\nabla h\cdot\mathbf{n})s\big]&\end{aligned}$$

where $n^2=n_1^2+n_2^2$. Hence, it follows that for s sufficiently large

$$-\nabla\cdot\big(h^3(x)\nabla\zeta\big)+g(x)\nabla h\cdot\nabla\zeta>0 \tag{3.10}$$

for all $x\in\Omega$. The inequalities (3.9) and (3.10) thus allow us to conclude that $(u_2-u_1)^+=0$ a.e. This concludes the proof of Theorem 3.

4. Nonnegativity and regularity of solutions. We have

$$\boldsymbol{\alpha}(x,0)=0,\qquad x\in\Omega. \tag{4.1}$$

Thus, Theorem 3 allows us to conclude that the weak solution, u, to (2.2) with boundary data $\varphi>0$ on $\partial\Omega$ is nonnegative, i.e., $u\geqq 0$ a.e. in Ω. (Take $\varphi_1=\varphi$, $u_1=u$, $\varphi_2=0$, $u_2=0$ in Theorem 3.) Thus, we have the following result.

COROLLARY 1. *If $\varphi>0$ a.e. in $\partial\Omega$, the weak solution, u, to* (2.2) *satisfies $u\geqq 0$ a.e. in Ω.*

If $\lambda\neq 0$ and $h(x)\in C^\infty(\bar\Omega)$, then $\boldsymbol{\alpha}(x,v)$ has bounded derivatives of all orders for $v\geqq 0$. Hence, the standard techniques for the analysis of the regularity of solutions to elliptic boundary value problems [2], [6] can be used to prove that u is a smooth, classical solution to (2.1). However, if $\lambda=0$, then $\boldsymbol{\alpha}(x,v)$ is not differentiable at $v=0$, and it is necessary to show that u is bounded away from zero to be able to prove that u is smooth. Thus, we demonstrate the following theorem which gives a condition on h which guarantees that u is bounded away from zero.

COROLLARY 2. *If $\varphi\geqq\Phi>0$ on $\partial\Omega$ where Φ is a constant and if*

$$\nabla\cdot\boldsymbol{\alpha}(x,\Phi)\leqq 0,\qquad x\in\Omega,$$

then the weak solution, u, to (2.2) *satisfies the bound $u\geqq\Phi>0$ on Ω.*

Proof. We will use the method of Theorem 3 with $\varphi_1=\varphi$, $u_1=u$, $\varphi_2=\Phi$, $u_2=\Phi$. Now $u_2=\Phi$ is not the weak solution of (2.2) with boundary data, $\varphi_2=\Phi$, but we have instead by integration by parts that for all $\zeta\in H_0^1(\Omega)$, $\zeta\geqq 0$,

$$\int_\Omega h^3(x)\nabla u_2\cdot\nabla\zeta-\boldsymbol{\alpha}(x,u_2)\cdot\nabla\zeta\,dx=\int_\Omega\nabla\cdot\boldsymbol{\alpha}(x,\Phi)\zeta\,dx\leqq 0. \tag{4.2}$$

From (4.2) we obtain in place of (3.3) that

$$\int_\Omega h^3(x)\nabla(u_2-u_1)\cdot\nabla\zeta\,dx-(\alpha(x,u_2)-\alpha(x,u_1))\cdot\nabla\zeta\,dx\leqq 0 \tag{4.3}$$

for all $\zeta\in H_0^1(\Omega)$, $\zeta\geqq 0$.

The proof of Corollary 2 now follows from the observation that the inequality (4.3) can replace the equality (3.3) in the argument of Theorem 3.

An argument similar to that given in Corollary 2 can be used to prove the following result.

COROLLARY 3. *If* $0\leqq\varphi\leqq\Phi$ *on* $\partial\Omega$ *and if*

$$\nabla\cdot\alpha(x,\Phi)\geqq 0,\qquad x\in\Omega,$$

then the weak solution, u, *to* (2.2) *satisfies the bound* $0\leqq u\leqq\Phi$ *on* Ω.

Note that the technique of Corollary 2 and Corollary 3 could provide some other comparison results by choosing φ_2 and u_2 to satisfy a suitable differential inequality. The bound given in Corollary 2 allows the standard regularity theory for elliptic partial differential equations [2], [6] to be used to prove the following theorem.

COROLLARY 4. *Assume that the boundary of* Ω, $\partial\Omega$, *is infinitely differentiable, that* $h\in C^\infty(\bar\Omega)$, *and that*

$$\Lambda\cdot\nabla h\leqq 0,\qquad x\in\Omega. \tag{4.4}$$

Then the Reynolds lubrication equation (1.1) *has a nonnegative classical solution,* P, *such that* $P\in C^\infty(\bar\Omega)$ *and such that* P *is unique in the class of nonnegative solutions.*

We note that it is proven in [5] under the hypothesis (4.4) that solutions to (1.1) are positive, smooth, and unique in the class of smooth solutions. Our results here prove that such a solution exists and that the stronger bound of Corollary 2 holds. Further, if (4.4) holds, the following result proves that our solution to (1.1) is unique in the class of weak solutions.

COROLLARY 5. *Assume that* P *is a weak solution to* (1.1) *in the sense that* $P\equiv P_a>0$ *on* $\partial\Omega$ *and*

$$\int_\Omega h^3P\nabla P\cdot\nabla\xi-Ph\Lambda\cdot\nabla\xi\,dx=0,\qquad \xi\in H_0^1(\Omega), \tag{4.5}$$

where $P, P^2\in H^1(\Omega)$. *If* (4.4) *holds, then* $P>0$ *a.e. in* Ω.

Proof. Take $\xi=P^-=\min(P,0)\in H_0^1(\Omega)$. Then by (4.5)

$$\begin{aligned}\int_\Omega h^3P\nabla P\cdot\nabla P^- - Ph\Lambda\cdot\nabla P^-\,dx &\\ =\int_\Omega h^3P^-|\nabla P^-|^2+\frac{1}{2}h\Lambda\cdot\nabla(P^-)^2\,dx &\\ =\int_\Omega h^3P^-|\nabla P^-|^2-\frac{1}{2}(\Lambda\cdot\nabla h)(P^-)^2\,dx=0.&\end{aligned} \tag{4.6}$$

Hence, by (4.4),

$$\int_\Omega h^3P^-|\nabla P^-|^2\,dx=\frac{1}{2}\int_\Omega(\Lambda\cdot\nabla h)(P^-)^2\,dx\leqq 0.$$

Thus, $P^-=0$ a.e. in Ω.

REFERENCES

[1] R. A. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.
[2] S. AGMON, *Elliptic Boundary Value Problems*, Van Nostrand, New York, 1965.
[3] A. BURGDORFER, *The influence of the molecular mean-free path on the performance of hydrodynamic gas lubricated bearings*, Trans. ASME, Part D, J. Basic Engrg., 81 (1959), pp. 94–100.
[4] J. CARILLO AND M. CHIPOT, *On some nonlinear elliptic equations involving derivatives of the nonlinearity*, Proc. Roy. Soc. Edinburgh, to appear.
[5] G. CIMATTI, *On certain nonlinear problems arising in the theory of lubrication*, Appl. Math., 11 (1984), pp. 227–245.
[6] D. GILBARG AND N. S. TRUDINGER, *Elliptic Partial Differential Equations of Second Order*, 2nd ed., Springer-Verlag, Berlin, 1981.
[7] W. GROSS, L. MATSCH, V. CASTELLI, A. ESHEL, J. VOHR AND M. WILDMANN, *Fluid Film Lubrication*, John Wiley, New York, 1980.
[8] B. WOLF, N. DESHPANDE, AND V. CASTELLI, *The flight of a flexible tape over a cylinder with a protruding bump*, AMSE J. Lubrication Technology, 105 (1983), pp. 138–142.

SIAM J. MATH. ANAL.
Vol. 17, No. 6, November 1986

ON THE RELATION BETWEEN COEFFICIENT AND BOUNDARY VALUES FOR SOLUTIONS OF WEBSTER'S HORN EQUATION*

WILLIAM W. SYMES†

Abstract. Webster's horn equation is a normalized version of the one-dimension linear acoustic wave equation. It has been used extensively as a simple model for plane wave propagation in layered systems, and particularly as the arena for much work on the relation between the acoustic impedance (coefficient) and the surface response or seismogram (boundary value) in theoretical seismology (this is the simplest so-called seismic reflection inverse problem). The question of continuous dependence of the solution on the coefficient arises naturally in this context, particularly in connection with perturbational techniques.

We study the dependence of boundary values for solutions of Webster's horn equation on its coefficient. For suitable choice of topologies (Sobolev spaces), we show that the map from coefficient to boundary values is a C^1-diffeomorphism.

Key words. wave propagation, one-dimensional acoustics, inverse problems, stability

AMS(MOS) subject classification. Primary 35R25

Introduction. Various wave propagation problems in a plane-stratified half-space $\{z>0\}$ lead to *Webster's horn equation*:

$$(\eta(z)\partial_t^2-\partial_z\eta(z)\partial_z)u(z,t)=0 \tag{0.1}$$

where the coefficient η is generally called an impedance, and is related to the mechanical properties of the medium (see [5] or [26] for a derivation and discussion of (0.1) in linear acoustics and elastodynamics).

We impose the boundary and initial conditions

$$\partial_z u(0,\cdot)=f, \tag{0.2}$$

$$u(\cdot,t)\equiv 0, \qquad t\ll 0, \tag{0.3}$$

where the Neumann datum f vanishes for $|t|$ large, so that the resulting disturbance is transient.

The conditions (0.1), (0.2), (0.3) form a well-posed initial boundary value problem in the half-space $(z>0)$, even for distribution datum f. Therefore, specification of η and f determines for instance the values of u on $\{z=0\}$. We define

$$\tau_f(\eta)=\partial_t u(0,\cdot). \tag{0.4}$$

Our aim in this paper is to establish the properties of τ_f as a map between suitable function spaces. Note that

$$\tau_f(\eta)=f*\tau_\delta(\eta).$$

Thus properties of τ_f follow from properties of τ_δ via well-known facts about convolution. Therefore, we restrict our attention to the choice $f=-\delta$ (the so-called impulse-response case). Our main result is that $\tau=\tau_{-\delta}$ is a Lipschitz homeomorphism, in the following sense.

*Received by the editors September 13, 1984.

†Department of Mathematics, Michigan State University, East Lansing, Michigan 48824. Present address, Department of Mathematical Sciences, Rice University, Houston, Texas 77001. This research was supported in part by the National Science Foundation under grant MCS-80-02996-01, and by the Office of Naval Research under contract N00014-83-K-0051.

We may rewrite (0.1) as

$$(0.1') \qquad \left(\partial_t^2 - \partial_z^2 - \sigma\partial_z\right)u = 0$$

with $\sigma = \partial_z \log \eta$ (the *reflectivity* in seismic parlance), and normalize $\eta(0) = 1$, by scaling u, so

$$\eta(z) = \exp\left\{\int_0^z \sigma\right\}.$$

We can view τ as a functional of σ instead, i.e. redefine

$$(0.4') \qquad \tau(\sigma) = \partial_t u(0, \cdot).$$

Note that the boundary value $u(o,t)$, $0 \leqq t \leqq 2T$, depends on the coefficient η (or σ) only in $0 \leqq z \leqq T$, since the signal propagation speed of (0.1) is one. The solution of (0.1), (0.2) and (0.3) is constructed by well-known methods (see below) for smooth η (or σ), so we can view τ initially as a map: $C^\infty[0,T] \to C^\infty[0,2T]$.

We can now state our three major results:

THEOREM 0.1. *For $\sigma \in C^\infty[0,T]$, there exist constants β and C^*, depending only on $\|\sigma\|_{L^2[0,T]}$ and on T, so that for $\tilde{\sigma} \in C^\infty[0,T]$ with $\|\sigma - \tilde{\sigma}\|_{L^2[0,T]} < \beta$ we have*

$$\|\tau(\sigma) - \tau(\tilde{\sigma})\|_{L^2[0,2T]} \leqq C^* \|\sigma - \sigma^*\|_{L^2[0,T]}.$$

So τ extends to a locally Lipschitz map:

$$L^2[0,T] \to L^2[0,2T].$$

THEOREM 0.2. *For $\sigma \in L^2[0,T]$, $g = \tau(\sigma) \in L^2[0,2T]$, there exist α and $C_* > 0$ depending only on $\|\sigma\|_{L^2[0,T]}$ and on T, for which*

$$\|\tilde{g} - g\|_{L^2[0,2T]} < \alpha$$

implies that $\tilde{g} = \tau(\tilde{\sigma})$ for a unique $\tilde{\sigma} \in L^2[0,T]$, with

$$\|\sigma - \tilde{\sigma}\|_{L^2[0,T]} < C_* \|g - \tilde{g}\|_{L^2[0,2T]}.$$

Actually Theorems 0.1 and 0.2 were the main results of our previous paper [21]. The arguments used in [21], however, suffered from two major drawbacks: they depend on

(i) a characterization of the range of τ; and

(ii) Lipschitz estimates for the time-like Cauchy problem.

Neither (i) nor (ii) is available in similar higher-dimensional problems. In the present treatment, we have reformulated the results so that the Lipschitz constants and domains of validity depend only on local quantities, thus avoiding (i). Also, we have revised the main step in the proof ("downward continuation") so that the time-like Cauchy problem no longer plays an explicit role. As a by-product of our analysis, we obtain a stronger result:

THEOREM 0.3. *τ is a C^1-diffeomorphism of $L^2[0,T]$ into $L^2[0,2T]$.*

As intended, the new approach has allowed us to establish carefully formulated analogues of Theorems 0.1, 0.2, and 0.3 for similar higher-dimensional problems: see [22] and references cited therein.

In the remainder of this introduction, we shall:

(i) explain the origin of our interest in the map τ;

(ii) briefly discuss related literature;

(iii) describe the main components of our analysis;

(iv) outline the organization of the paper.

Our interest in the properties of τ_f is motivated by the *impedance profile inversion problem*, which is a model for many inverse (or coefficient—or medium—identification) problems in wave propagation. In the impedance profile inversion problem, a propagating plane-wave disturbance (measured by u) is introduced into a medium $\{z>0\}$ (modeled by (0.1)) by means of a boundary excitation (imposed traction, pressure drop, etc., modeled by (0.2)), assumed quiescent in the past ((0.3)). The inhomogeneities in the medium ($\eta \neq$ constant) set up reflected waves, which reach the surface and result in a nonzero rate of change of the boundary value of u ($\tau_f(\sigma) = \partial_t u(0,\cdot)$). From this latter data, one is to infer the structure of the medium (i.e. $\eta(z)$, $z>0$).

In our notation, we formulate the impedance profile inversion problem as follows: given data g and f, find η (or $\sigma = \partial_z \log \eta$) so that the functional equation

$$\tau_f(\sigma) = g \tag{0.5}$$

is satisfied.

This problem and its generalizations are models for many interesting data-pocessing technologies, the best-known being the seismic reflection method in petroleum prospecting. Correspondingly, inverse problems in wave propagation have a large and rapidly growing literature, much of it concerned with impedance profile inversion and simple variants (layered-medium problems). Therefore, it seems surprising that the properties of the map τ_f have seldom been addressed in any substantial way, even though many authors have recommended perturbation or optimization techniques for (0.5) and its generalizations, which require properties such as differentiability (see [7], [13], and references cited there). Even continuity of τ and its inverse has been addressed in only a few papers (see [21], [6], [11], [1], and in the context of spectral inverse problems [14]). The issue is serious, as any measurement modeled by τ is inevitably contaminated with noise, and τ and τ^{-1} are *not* continuous with respect to some apparently natural domain and range metrics (see [11] for discontinuity examples in impedance profile inversion, and [23] for different examples of nondifferentiability of τ and τ^{-1} in higher-dimensional problems). Moreover, many approaches to impedance profile inversion do not generalize in any obvious way to higher-dimensional (nonlayered) problems.

Our approach in this paper is a modification of the *downward continuation* (or "layered-stripping") method, which really goes back to the geophysical work of Goupillaud [12] and Kunetz [15] in the early 1960s. A number of authors recently have developed reliable computational techniques for impedance profile inversion based on this idea ([17], [4], [3], [19], [9], [2], [20]).

The main tools used in the present study (and in [21]) are:

(I) *The transport equation*. The progressing wave expansion ([8, Chap. VI, §4, esp. pp. 633–655], also [23, §2], and [10, pp. 42–45]) for smooth η shows that the solution of (0.1), (0.2), (0.3) with $f = -\delta$ must have a jump along the characteristic $\{z=t\}$ of magnitude

$$[u](z,z) = \eta^{-1/2}(z)$$

(with the normalization $\eta(0)=1$), and is otherwise smooth. Therefore, u solves the characteristic initial boundary value problem

$$(\partial_t^2 - \partial_z^2 - \sigma\partial_z)u = 0 \quad \text{in } \Omega := \{(z,t): 0<z<T,\ z<t<2T-z\}, \tag{0.6a}$$

$$\partial_z u(0,t) = 0, \qquad 0<t<2T, \tag{0.6b}$$

$$\lim_{\zeta \to 0^+} u(z, z+\zeta) =: \bar{u}(z) = \eta^{-1/2}(z) \tag{0.6c}$$

with

$$\eta(z)=\exp\int_0^z\sigma. \tag{0.6d}$$

In effect the use of (0.6) rather than (0.1), (0.2), (0.3) removes the propagating singularity from the problem.

(II) *Downward continuation problem.* We shall actually study a slightly more general problem than (0.6), because the generalization does not significantly increase the difficulty of the arguments and because the extended results are useful in further investigation of the inverse problem as an optimization problem with constraints on η (or σ).

We note that, if we pick $0\leqq z_0<z_1\leqq T$ and set

$$f:=\partial_z u(z_0,\cdot), \tag{0.7a}$$

$$h:=(\partial_z+\partial_t)u(z_1,\cdot), \tag{0.7b}$$

then in the region $\{(z,t): z_0\leqq z\leqq z_1,\ z\leqq t\leqq 2T-z\}$ the solution u of (0.6) satisfies

$$(\partial_t^2-\partial_z^2-\sigma\partial_z)u=0, \tag{0.8a}$$

$$\partial_z u(z_0,\cdot)=f, \tag{0.8b}$$

$$(\partial_t+\partial_z)u(z_1,\cdot)=h, \tag{0.8c}$$

$$\bar{u}(z)=\eta^{-1/2}(z). \tag{0.8d}$$

Now define, for arbitrary f, h, σ, η_0

$$\tau_0(\sigma,h,f,\eta_0):=\partial_t w(z_0,\cdot)$$

where

$$\eta(z)=\eta_0\exp\left\{\int_{z_0}^z\sigma\right\}$$

and w solves the mixed characteristic initial-boundary value problem (0.8).

We shall show in §2 that, if we regard f and η_0 as parameters, the equation

$$\tau_0(\sigma,h,f,\eta_0)=g \tag{0.9}$$

has at most one solution (σ,h) which depends continuously on f, η_0 and g. In fact, we shall show that τ_0 is Lipschitz in all variables, continuously differentiable in the first two, and we shall bound the derivative and its inverse locally. The implicit function theorem and a global uniqueness theorem then yield the result.

Note also that if we pick $z_0=0$, $z_1=T$, then $\tau_0(\sigma,0,0,1)=\tau(\sigma)$. Thus Theorems 0.1–0.3 follow from similar statements about τ_0.

The uniqueness of solutions to the functional equation for τ_0 also implies that if f is given by (0.7a), $g=\partial_t u(z_0,\cdot)$, and $\eta_0=\eta(z_0)$ from (0.6d), then the solution of (0.8) is identical to the solution of (0.6) for $z_0\leqq z\leqq z_1$. In particular, the solution (σ,h) to (0.9) *continues* σ to $[z_0,z_1]$. Also, in solving (0.8), we have implicitly constructed $\partial_t u(z_1,\cdot)$ and $\partial_z u(z_1,\cdot)$. So we can redefine g, f, and η_0, replace z_0 by z_1, and continue the procedure. This is a version of the downward continuation algorithm suited to several-dimensional generalizations.

Most inverse problems of wave propagation which model feasible experiments are ill-posed, either because of explicit band-limitation of the observed signal or because of several-dimensional effects (see [22], [23] for a discussion of the latter). This is even true

for some versions of the impedance problem with band limited data (the problem considered in this paper appears to be the sole exception). Such problems must be set as regularized (nonlinear) optimization problems, and the presence of multiple local minima becomes a possibility.

The global minimum of $\|\tau_0 - g\|$, however, is unique in a region of size proportional to $(z_1 - z_0)^{-1/2}$. This fact may be used to approximate recursively the global minimum of $\|\tau - g\|$ by a nonlinear least-squares version of downward continuation. This approach, which is a particular homotopy method for global optimization of $\|\tau - g\|$, underlies our interest in the map τ_0, and will be discussed elsewhere.

(III) *Sideways energy estimates.* The key tool in constructing derivatives and bounds for τ_0 is the estimation of

$$Q(z) := \frac{1}{2}\int_z^{2T-z} dt \left\{ (\partial_t u)^2 + (\partial_z u)^2 \right\}$$

for solutions u of (0.8). Q is a sort of vertical energy form. Its interaction with the usual energy form produces most of the important estimates. The key point is that the $1-D$ wave equation is hyperbolic as an evolution equation in z as well. This sort of argument seems to have been introduced by Rauch and Taylor [16], who used it to prove energy decay for some dissipative boundary value problems in $1+1$ dimensions. Of course, the time-like Cauchy problem is no longer hyperbolic in $1+n$ dimensions, $n>1$, so the argument breaks down, as noted in [16]. The author has recently studied regularization of the higher-dimensional time-like Cauchy problem, in such a way that "sideways" energy estimate for analogues of Q are again useful (see [22], [25]).

The paper is organized as follows:

Section 1: derivation of a number of "sideways" energy estimates.

Section 2: investigation of the downward continuation operator τ_0, proof of Theorems 0.1–0.3.

In conclusion, we note that any serious attempt to model applications such as reflection seismology by the impedance profile inversion problem sketched above runs into serious, and as yet mostly unresolved, difficulties. The gravest of these concerns the Neumann datum f, which for any realistic model is far from impulsive, and in fact tends to have very small Fourier components near both 0 and ∞ *Hz*. For the disastrous effect of the low-frequency small amplitudes on coefficient determination, and a suggestion for a partial cure, see [18]. See also, however, reference [20] in which these measurement difficulties are overcome and the solution of the impedance profile inversion problem by numerical downward continuation allows the real-time imaging of the vocal tract.

1. Energy estimates. Suppose $0 \leqq z_0 < z_1 \leqq T$, and define $\Omega = \{(z,t) : z_0 \leqq z \leqq z_1,\ z \leqq t \leqq 2T - z\}$. Thus Ω is the trapezoid obtained by intersection the double light cone, with vertices $(0,0)$, $(0,2T)$ with the strip $\{(z,t) : z_0 \leqq z \leqq z_1\}$. Suppose $u, w \in C^\infty(R^2)$, $\eta_0 \in R^+$, ω, η, σ, f, h and $\psi \in C^\infty(R)$ satisfy

$$\eta(z) = \eta_0 \exp \int_{z_0}^{z} \sigma, \qquad z_0 \leqq z \leqq z_1, \tag{1.1}$$

$$\left(\partial_t^2 - \partial_z^2 - \sigma\partial_z\right) u = \omega w \quad \text{in } \Omega, \tag{1.2a}$$

$$\partial_z u(z_0, t) = f(t), \qquad z_0 \leqq t \leqq 2T - z_0, \tag{1.2b}$$

$$(\partial_z + \partial_t) u(z_1, t) = h(t), \qquad z_1 \leqq t \leqq 2T - z_1, \tag{1.2c}$$

$$\bar{u}(z) = \psi(z), \qquad z_0 \leqq z \leqq z_1. \tag{1.2d}$$

Note: We will make use throughout of the symbol $\bar{u}$ to represent the restriction of a real-valued function of two variables to the diagonal $z=t$. We will also use the abbreviation

$$\bar{\bar{u}}(z) := u(z, 2T-z)$$

to indicate restriction to the upper boundary segment of Ω.

The problem (1.2) may be viewed as a mixed characteristic initial/boundary value problem with data σ, ω, w, f, h. A standard construction ([7], pp. 461–471) assures the existence of a solution u depending continuously on the data in various senses. The purpose of this section is to establish estimates for the boundary values of u and its restrictions to vertical line segments.

The main tools are two "energy identities." These are most easily derived by application of Stoke's theorem. The first is related to the well-known energy method ([8, pp. 438–449]); set

$$\omega_E = \frac{1}{2}\eta\left\{(\partial_t u)^2 + (\partial_z u)^2\right\} dz + \eta \partial_z u \partial_t u \, dt;$$

Stokes' theorem

$$\int_{\partial\Omega} \omega_E = \int_{\Omega} d\omega_E$$

amounts to the identity

$$\begin{aligned}(1.3)\qquad \frac{1}{2}\int_{z_0}^{z_1} \eta(\bar{u}')^2 - \frac{1}{2}\int_{z_0}^{z_1} \eta(\bar{\bar{u}}')^2 + \eta(z_1)\int_{z_1}^{2T-z_1} dt\, \partial_t u \partial_z u(z_1, t) \\ - \eta(z_0)\int_{z_0}^{2T-z_0} dt\, \partial_z u \partial_t u(z_0, t) \\ = \int_{z_0}^{z_1} dz\, \eta(z)\omega(z)\int_{z}^{2T-z} dt\, w(z,t)\partial_t u(z,t)\end{aligned}$$

where in rewriting the area integral we have made use of the form of (1.2a):

$$\left(\eta\partial_t^2 - \partial_z \eta \partial_z\right) u = \eta\omega w.$$

The second identity involves the "sideways energy" form:

$$\omega_Q = \tfrac{1}{2}\left\{(\partial_t u)^2 + \left(\partial_z u^2\right)\right\} dt + \partial_t u \partial_z u \, dz.$$

We introduce the abbreviations

$$\begin{aligned} Q(z) &:= \frac{1}{2}\int_z^{2T-z} dt \left\{(\partial_t u)^2 + (\partial_z u)^2\right\}(z,t), \\ \Omega_{a,b} &= \left\{(z,t) : z_a \leqq z \leqq z_b,\ z \leqq t \leqq 2T-z\right\} \end{aligned}$$

for $z_0 \leqq z_a < z_b \leqq z_1$. Then the identity

$$\int_{\partial\Omega_{a,b}} \omega_Q = \int_{\Omega_{a,b}} d\omega_Q$$

amounts to

$$Q(z_b)-Q(z_a)+\frac{1}{2}\int_{z_a}^{z_b}\{(\bar{u}')^2+(\bar{\bar{u}}')^2\} \tag{1.4}$$
$$=-\int_{z_a}^{z_b}dz\,\sigma(z)\int_z^{2T-z}dt(\partial_z u(z,t))^2$$
$$-\int_{z_a}^{z_b}dz\,\omega(z)\int_z^{2T-z}dt\,w(z,t)\partial_z u(z,t).$$

The identities (1.3) and (1.4) together imply a number of useful estimates on boundary values of u, corresponding to various special cases of (1.2). We make the convention in the following lemmas and their proofs that constants, which are denoted by C and which change their meaning from equation to equation, depend only on $\|\sigma\|$, η_0, T, and z_1-z_0, unless otherwise noted, and that $\|\cdot\|$ refers to the L^2-norm on the appropriate interval.

LEMMA 1.1. *Suppose in* (1.2) *that* $h=\omega\equiv 0$. *Then*

$$Q(z)\leqq C\{\|\psi'\|^2+\|f\|^2\}$$

for $z_0\leqq z\leqq z_1$.

Proof. Setting $\omega=0$ and $z_a=z$, $z_b=z_1$ in (1.4), we obtain

$$Q(z_1)+\frac{1}{2}\int_z^{z_1}(\psi')^2+\frac{1}{2}\int_z^{z_1}(\bar{\bar{u}}')^2=Q(z)-\int_z^{z_1}dz\,\sigma(z)\int_\tau^{2T-z}dt(\partial_z u)^2(\zeta,t).$$

Thus,

$$Q(z)\leqq\frac{1}{2}\int_z^{z_1}|\psi'|^2+\left\{Q(z_1)+\frac{1}{2}\int_z^{z_1}(\bar{\bar{u}}')^2\right\}+2\int_z^{z_1}d\zeta|\sigma(\zeta)|Q(\zeta).$$

So Gronwall's inequality implies

$$Q\leqq\left\{\frac{1}{2}\|\psi'\|^2+\left[Q(z_1)+\frac{1}{2}\|\bar{\bar{u}}'\|^2\right]\right\}\exp\left(2\int_{z_0}^{z_1}|\sigma|\right). \tag{1.5}$$

Since $(\partial_t+\partial_z)u(z_1,\cdot)=0$, we have

$$\partial_t u\partial_z u=-\tfrac{1}{2}\left[(\partial_t u)^2+(\partial_z u)^2\right]\quad\text{for } z=z_1.$$

Therefore (1.3) yields

$$\frac{1}{2}\int_{z_0}^{z_1}\eta(\psi')^2-\eta(z_0)\int_{z_0}^{2T-z_0}dtf(t)\partial_t u(z_0,t)=\frac{1}{2}\int_{z_0}^{z_1}\eta(\bar{\bar{u}}')^2+\eta(z_1)Q(z_1) \tag{1.6}$$
$$\geqq\eta_*\left(\frac{1}{2}\|\bar{\bar{u}}'\|^2+Q(z_1)\right)$$

where

$$\eta^*=\sup\eta(z)\leqq\eta_0\exp\int_{z_0}^{z_1}|\sigma|\leqq\eta_0\exp\{(z_1-z_0)^{1/2}\|\sigma\|\},$$
$$\eta_*=\inf\eta(z)\geqq\eta_0\exp\{-(z_1-z_0)^{1/2}\|\sigma\|\}.$$

The L.H.S. of (1.6) is bounded by

$$\eta^*\left(\frac{1}{2}\|\psi'\|^2+\|f\|Q(z_0)^{1/2}\right)\leqq\eta^*\left(\frac{1}{2}\|\psi'\|^2+\frac{k}{2}\|f\|^2+\frac{1}{2k}Q(z_0)\right)$$

for any $k>0$. Combined with (1.5) this yields

$$Q\leqq C\left\{\|\psi'\|^2+kC_2\|f\|^2+\frac{\eta^*}{2\eta_* k}Q(z_0)\right\}\exp 2\int_{z_0}^{z_1}|\sigma|.$$

Take $k=(\eta^*/\eta_*)\exp\{2\int_{z_0}^{z_1}|\sigma|\}$; then we have

$$Q\leqq C\left\{\|\psi'\|^2+\|f\|^2+\tfrac{1}{2}Q(z_0)\right\}$$

so

$$2Q(z_0)\leqq 2C\left\{\|\psi'\|^2+\|f\|^2\right\}$$

and so

$$Q\leqq 2C\left\{\|\psi'\|^2+\|f\|^2\right\}. \qquad \text{Q.E.D.}$$

LEMMA 1.2. *Suppose in* (1.2) *that* $f=h=\psi=0$. *Then,*

$$Q\leqq C(z_1-z_0)\|\omega\|^2 w^*$$

where

$$w^*=\sup\left(\int_z^{2T-z}dt w^2(z,t)\right).$$

Proof. From (1.4) we obtain

$$Q(z_1)=Q(z)-\frac{1}{2}\int_z^{z_1}(\bar{\bar{u}}')^2-\int_z^{z_1}d\zeta\sigma(\zeta)\int_\zeta^{2T-\zeta}dt(\partial_z u)^2(\zeta,t)$$
$$-\int_z^{z_1}d\zeta\omega(\zeta)\int_\zeta^{2T-\zeta}dt(w\partial_z u)(\zeta,t).$$

So

$$(1.7)\quad Q(z)\leqq Q(z_1)+\frac{1}{2}\int_{z_0}^{z_1}(\bar{\bar{u}}')^2+\int_z^{z_1}d\zeta\int_\zeta^{2T-\zeta}dt\left\{|\sigma(\zeta)|(\partial_z u)^2(\zeta,t)\right.$$
$$\left.+\|\omega(\zeta)\|w(\zeta,t)\,d_z u|\zeta,t|\right\}.$$

Identity (1.3) gives

$$-\frac{1}{2}\int_{z_0}^{z_1}\eta(\bar{\bar{u}}')^2+\eta(z_1)\int_{z_1}^{2T-z_1}dt\partial_t u\partial_z u(z_1,t)$$
$$=\int_{z_0}^{z_1}dz\eta(z)\omega(z)\int_z^{2T-z}dtw(z,t)\partial_t u(z,t).$$

As in the proof of Lemma 1.1, $\partial_t u\partial_z u=-\frac{1}{2}[(\partial_t u)^2+(\partial_z u)^2]$ at $z=z_1$, so the above implies

$$(1.8)\qquad \eta_*\left(\frac{1}{2}\int_{z_0}^{z_1}(\bar{\bar{u}}')^2+Q(z_1)\right)\leqq\eta^*\int_{z_0}^{z_1}d\zeta|\omega(\zeta)|\int_\zeta^{2T-\zeta}dt|w\partial_\tau u|(\zeta,T).$$

Now for any $\alpha > 0$

$$\int dt|w||\partial_t u| = \int dt(\alpha|w|)\left(\frac{1}{\alpha}|\partial_t u|\right) \leqq \frac{1}{2}\int dt\left(\alpha^2|w|^2 + \alpha^{-2}|\partial_t u|^2\right) \tag{1.9}$$

$$\leqq \frac{\alpha^2 w^*}{2} + \frac{\alpha^{-2}}{2}\int |\partial_t u|^2.$$

Similarly

$$\int dt|w||\partial_z u| \leqq \frac{\alpha^2 w^*}{2} + \frac{\alpha^{-2}}{2}\int y\,|\partial_z u|^2. \tag{1.10}$$

Equations (1.8), (1.9), and (1.10) combine with (1.7) to yield

$$Q(z) \leqq \int_{z_0}^{z_1} d\zeta|\omega(\zeta)|\left(1+\frac{\eta^*}{\eta_*}\right)(\alpha^2 w^* + 2\alpha^{-2}Q^*) + 2\int_{z_0}^{z_1}|\sigma|Q$$

where $Q^* := \sup_{z_0 \leqq z \leqq z_1} Q(z)$. By Gronwall,

$$Q \leqq C(\alpha^2 w^* + 2\alpha^{-2}Q^*)\left\{\int_{z_0}^{z_1}|\omega|\right\}\exp\left\{2\int_{z_0}^{z_1}|\sigma|\right\}. \tag{1.11}$$

Now if $\omega = 0$ a.e., $u \equiv 0$ by uniqueness in the mixed problem, and the assertion of the lemma holds trivially. Otherwise, set

$$\alpha^2 = 4\int_{z_0}^{z_1}|\omega|\left(\exp 2\int_{z_0}^{z_1}|\sigma|\right)$$

so that (1.11) becomes

$$Q \leqq C\left(\int_{z_0}^{z_1}|\omega|\right)^2 w^* \exp\left(4\int_{z_0}^{z_1}|\sigma|\right) + \frac{1}{2}Q^* \leqq C(z_1 - z_0)w^*\|\omega\|^2 + \frac{1}{2}Q^*$$

from which the conclusion follows immediately. Q.E.D.

LEMMA 1.3. *Suppose in* (1.2) *that* $f = \omega = \psi = 0$. *Then*

$$Q \leqq C\|h\|^2.$$

Proof. From

$$(\partial_t + \partial_z)u(z_1, \cdot) = h$$

we obtain

$$\int_{z_1}^{2T-z_1} dt\,\partial_t u \partial_z u(z_1, t) = \frac{1}{2}\int_{z_1}^{2T-z_1} h^2 - Q(z_1).$$

So the identity (1.3) reads

$$0 = \frac{1}{2}\int_{z_0}^{z_1}\eta(\bar{\bar{u}}')^2 + \eta(z_1)\left(\frac{1}{2}\int_{z_1}^{2T-z_1}h^2 - Q(z_1)\right)$$

whence

$$\frac{1}{2}\int_{z_1}^{2T-z_1}h^2 = \frac{1}{2}\int_{z_0}^{z_1}\eta(\bar{\bar{u}}')^2 + \eta(z_1)Q(z_1) \geqq \eta_*\left(\frac{1}{2}\int_{z_0}^{z_1}(\bar{\bar{u}}')^2 + Q(z_1)\right). \tag{1.12}$$

On the other hand, (1.4) gives

$$Q(z)+\int_z^{z_1} d\zeta\sigma(\zeta)\int_\zeta^{2T-\zeta} dt(\partial_z u)^2(\zeta,t)=Q(z_1)+\frac{1}{2}\int(\bar{\bar{u}}')^2. \tag{1.13}$$

Equations (1.12) and (1.13) combine to yield

$$Q(z)\leqq C\|h\|^2+\int_z^{z_1}|\sigma|Q$$

whence the result follows via Gronwall's inequality. Q.E.D.

LEMMA 1.4. *The solution of* (1.2) *satisfies*

$$Q(z)\leqq C\left\{\|\psi'\|^2+\|f\|^2+\|h\|^2+w^*\|\omega\|^2\right\}$$

in the notation of Lemmas 1.1–1.3.

Proof. This follows immediately from the previous lemmas by writing

$$u=u_1+u_2+u_3$$

and correspondingly

$$Q\leqq 3\{Q_1+Q_2+Q_3\}$$

where u_k solves the problem of Lemma 1. k, $k=1,2,3$. Q.E.D.

LEMMA 1.5. *Suppose that there exists* $k>0$ *so that for* $z_0\leqq z,\ \leqq z_1$,

$$\int_{z_0}^z|\omega|^2\leqq k\left\{\int_{z_0}^z|\psi'|^2+|\psi(0)|^2\right\}.$$

Then there exists $C>0$ *so that*

$$\|\psi'\|^2+\|h\|^2\leqq C\left\{Q(z_0)+|\psi(0)|^2\right\}.$$

Proof. From (1.4) we obtain

$$\begin{aligned}
&Q(z)+\frac{1}{2}\int_{z_0}^z(\psi')^2\\
&\leqq Q(z_0)+\int_{z_0}^z d\zeta\left\{|\sigma(\zeta)|\int_\zeta^{2T-\zeta}dt\left(\partial_z w^2(\zeta,t)\right)+\left|\omega(\zeta)\int_\zeta^{2T-\zeta}dt\,w(\zeta,t)\partial_z u(\zeta,t)\right|\right\}\\
&\leqq Q(z_0)+\int_{z_0}^z|\sigma|Q+w^*\int_{z_0}^z d\zeta|\omega(\zeta)|\left(\int_\zeta^{2T-\zeta}dt(\partial_z u)^2(\zeta,t)\right)^{1/2}\\
&\leqq Q(z_0)+\delta w^*\int_{z_0}^z|\omega|^2+\int_{z_0}^z\left(\sigma+\frac{w^*}{4\delta}\right)Q
\end{aligned}$$

for any $\delta<0$. For δ sufficiently small, the hypothesis allows us to dominate the second term on the R.H.S. by a fraction of the second term on the L.H.S., after adding a suitable multiple of $|\psi(0)|^2$, to both sides. We replace Q by $Q(z)+\frac{1}{2}\int_{z_0}^z|\psi'|^2$ on the R.H.S. under the integral sign and apply Gronwall's inequality to obtain

$$Q(z)+\frac{1}{2}\int_{z_0}^z|\psi'|^2\leqq\left\{Q(z_0)+c|\psi(0)|^2\right\}\exp\left\{\int_{z_0}^z\left(\sigma+\frac{w^*}{4\delta}\right)\right\}.$$

The proof is finished by setting $z=z_1$ and noting that

$$\int_{z_1}^{2T-z_1}|h|^2 \leqq 4Q(z_1). \qquad \text{Q.E.D.}$$

2. Proofs of the main theorems. Suppose $T>0$, $0\leqq z_0<z_1<T$, $\eta_0>0$, $\sigma\in C^\infty[z_0,z_1]$, $f\in C^\infty[z_0,2T-z_0]$, $h\in C^\infty[z_1,2T-z_1]$. Define $\eta=\eta[\sigma,\eta_0]$ by

$$\eta(z)=\eta_0\exp\int_{z_0}^{z}\sigma.$$

Thus, $\sigma=\partial_z\log\eta$ and $\eta(z_0)=h_0$. Suppose u solves the boundary value problem

(2.1a) $\quad(\partial_t^2-\partial_z^2-\sigma\partial_z)u=0 \quad$ in $\Omega:=\{(z,t): z_0\leqq z\leqq z_1,\ z\leqq t\leqq 2T-z\}$,

(2.1b) $\quad\partial_z u(z_0,t)=f(t), \qquad z_0\leqq t\leqq 2T-z_0$,

(2.1c) $\quad(\partial_t+\partial_z)u(z_1,t)=h(t), \qquad z_1\leqq t\leqq 2T-z_1$,

(2.1d) $\quad\bar{u}(z)=\eta^{-1/2}(z), \qquad z_0\leqq z\leqq z_1$.

Define

$$\tau_0(\sigma,h,f,\eta_0)(t):=\partial_t u(z_0,t), \qquad z_0\leqq t\leqq 2T-z_0.$$

τ_0 is the basic downward continuation operator. It is well-defined, as the problem (2.1) has unique solutions depending continuously on the data, by standard reasoning (see [8, Chap. V, §6]).

Our first task is the extent τ_0 continuously to square-integrable arguments.

PROPOSITION 2.1. *Suppose σ, η, f, h, η_0, u are smooth and satisfy the equations* (2.1). *Then*

$$Q(z):=\frac{1}{2}\int_z^{2T-z}dt\left\{(\partial_t u)^2+(\partial_z u)^2\right\}(z,t) \tag{2.2}$$

satisfies

$$Q(z)\leqq C\left\{\|\sigma\|^2+\|h\|^2+\|f\|^2\right\}, \qquad z_0\leqq z\leqq z_1$$

where C is exponential polynomial in η_0, $\|\sigma\|$, and z_1-z_0.

Proof. Since

$$\left\|\partial_z\eta^{-1/2}\right\|\leqq C(\eta_0,\|\sigma\|,z_1-z_0)\|\sigma\|$$

use of (2.1d) and Lemma 1.4 imply this estimate. Q.E.D.

PROPOSITION 2.2. *Suppose u_i, h_i, σ_i, η_{0i}, η_i, and f_i are smooth and satisfy* (2.1), *$i=1,2$. Let $v=u_2-u_1$ and set*

$$Q(z):=\frac{1}{2}\int_z^{2T-z}dt\left\{(\partial_t v)^2+(\partial_z v)^2\right\}(z,t);$$

then for $z_0\leqq z\leqq z_1$,

$$Q(z)\leqq C\left\{\|\sigma_1-\sigma_2\|^2+\|h_1-h_2\|^2+\|f_1-f_2\|^2+|\eta_{01}-\eta_{02}|^2\right\}$$

where C is exponential/polynomial in

$$\eta_{0i},\quad \|\sigma_i\|,\quad \|f_i\|,\quad \|h_i\|,\quad z_1-z_0, \qquad i=1,2.$$

Proof. The difference v solves the boundary value problem

$$(\partial_t^2-\partial_z^2-\sigma_1\partial_z)v=(\sigma_1-\sigma_2)\partial_z u_2 \quad \text{in } \Omega, \tag{2.3a}$$

$$\partial_z v(z_0,t)=f_2(t)-f_1(t), \qquad z_0\leqq t\leqq 2T-z_0, \tag{2.3b}$$

$$(\partial_t+\partial_z)v(z_1,t)=h_2(t)-h_1(t), \qquad z_1\leqq t\leqq 2T-z_1, \tag{2.3c}$$

$$\bar{v}(z)=\eta_2^{-1/2}(z)-\eta_1^{-1/2}(z), \qquad z_0\leqq z\leqq z_1. \tag{2.3d}$$

Note that

$$\begin{aligned}\eta_2^{-1/2}(z)-\eta_1^{-1/2}&=n_{02}^{-1/2}\left(\exp\left(-\frac{1}{2}\int_{z_0}^{z}\sigma_2\right)-\exp\left(-\frac{1}{2}\int_{z_0}^{z}\sigma_1\right)\right)\\&\quad+\left(\eta_{02}^{-1/2}-\eta_{01}^{-1/2}\right)\left(\exp\left(-\frac{1}{2}\int_{z_0}^{z}\sigma_1\right)\right)\\&=\eta_2^{-1/2}\left[\left(1-\exp\left(\frac{1}{2}\int_{z_0}^{z}(\sigma_2-\sigma_1)\right)\right)+\left(1-\eta_{02}^{1/2}\eta_{01}^{-1/2}\right)\right]\end{aligned}$$

and $\|\partial_z\eta_i^{-1/2}\|\leqq C\|\sigma_i\|$. Thus, after some manipulation,

$$\left\|\partial_z\left(\eta_2^{-1/2}-\eta_1^{-1/2}\right)\right\|\leqq C\left\{\|\sigma_2-\sigma_1\|^2+|\eta_{02}-\eta_{01}|^2\right\}^{1/2}.$$

Also Proposition 2.1 shows that

$$\int_z^{2T-z}|\partial_z u_2|^2\leqq Q(z)\leqq C(\sigma_2,\,\eta_{20},\,f_2)$$

so the conclusion follows from Lemma 1.4 with $w\to\partial_z u_2$, $\omega\to\sigma_1-\sigma_2$, $h\to h_2-h_1$, $\psi\to\eta_2^{-1/2}-\eta_1^{-1/2}$. Q.E.D.

COROLLARY 2.3. *τ_0 extends to a locally Lipschitz map*:

$$L^2[z_0,z_1]\oplus L^2[z_1,2T-z_1]_-\oplus L^2[z_0,2T-z_0]\oplus R^+\to L^2[z_0,2T\text{-}z_0].$$

More precisely, for each (σ,h,f,η_0) there exist positive constants r and k for which

$$\|\sigma-\tilde{\sigma}\|^2+\|h-\tilde{h}\|^2+\|f-\tilde{f}\|^2+|\eta_0-\tilde{\eta}_0|^2<r^2$$

implies

$$\|\tau_0(\sigma,h,f,\eta_0)-\tau_0(\tilde{\sigma},\tilde{h},\tilde{f},\tilde{\eta}_0)\|<k\left(\|\sigma-\tilde{\sigma}\|^2+\|h-\tilde{h}\|^2+\|f-\tilde{f}\|^2+|\eta_0-\tilde{\eta}_0|^2\right)^{1/2}.$$

The parameters r and k are (bounded by) functions of $\|\sigma\|$, $\|f\|$, z_1-z_0, and η_0.

Proof of Theorem 0.1. If we take $z_0=0$, $z_1=T$, the dependence of τ_0 on the second argument disappears altogether. In fact, from the definitions in the introduction,

$$\tau(\sigma)=\tau_0(\sigma,0,0,1).$$

Thus Theorem 0.1 follows immediately from Corollary 2.3. Q.E.D.

For the sake of brevity set

$$K_D := L^2[z_0,z_1]\oplus L^2[z_1,2T-z_1]\oplus L^2[z_0,2T-z_0]\oplus R^+,$$
$$K_D^\infty := C^\infty[z_0,z_1]\oplus C^\infty[z_1,2T-z_1]\oplus C^\infty[z_0,2T-z_0]\oplus R^+,$$
$$H_D := L^2[z_0,z_1]\oplus L^2[z_1,2T-z_1],$$
$$H_T := L^2[z_0,2T-z_0].$$

The formal linearization of τ_0 in (σ,h) is given by

$$D\tau_0(\sigma,h,f,\eta_0)\cdot(\sigma^1,h^1) = \partial_t u^1\big|_{z=z_0}$$

where u solves (2.1) and u^1 solves

$$(\partial_t^2-\partial_z^2-\sigma\partial_z)u^1=\sigma^1\partial_z u \quad \text{in } \Omega, \tag{2.4a}$$
$$\partial_z u^1(z_0,\cdot)=0, \tag{2.4b}$$
$$(\partial_t+\partial_z)u^1(z_1,\cdot)=h^1, \tag{2.4c}$$
$$\bar{u}^1(z)=-\frac{1}{2}\eta^{-1/2}\int_{z_0}^z\sigma^1. \tag{2.4d}$$

All of this makes sense a priori when σ, h, f, σ^1, and h^1 are smooth. The first task is to extend $D\tau_0$.

THEOREM 2.4. *For smooth* $(\sigma,h,f,\eta_0)\in K_D^\infty$, $D\tau_0$ *extends to a bounded linear map:* $H_D\to H_T$. *Moreover,*

$$\|\sigma^1\|^2+\|h^1\|^2\leqq C\|D\tau_0(\sigma,h,f,\eta_0)\cdot(\sigma^1,h^1)\|^2 \tag{2.5}$$

so in particular the range of $D\tau_0$ *is closed in* H_T.

Proof. The boundedness of $D\tau_0$, hence the existence of the extension, follow immediately from the application of Lemma 1.4 to the boundary value problem (2.4).

A simple "Poincaré inequality" argument gives

$$\int_{z_0}^z|\sigma^1|^2\leqq C\int_{z_0}^z|\partial_z\bar{u}^1|^2, \qquad z_0\leqq z\leqq z_1$$

(or see [21, inequality (13)]). Then Lemma 1.5 yields the estimate (2.5). Q.E.D.

We next study the adjoint of $D\tau_0$. Suppose that v solves the boundary value problem

$$(\eta\partial_t^2-\partial_z\eta\partial_z)v=0 \quad \text{in } \Omega, \tag{2.6a}$$
$$\partial_z v(z_0,\cdot)=\eta_0^{-1}\phi\in C^\infty[z_0,2T-z_0], \qquad \phi(2T-z_0)=0, \tag{2.6b}$$
$$(\partial_t-\partial_z)v(z_1,\cdot)=0, \tag{2.6c}$$
$$\bar{\bar{v}}\equiv 0. \tag{2.6d}$$

Then Green's theorem, i.e. integration by parts of

$$\int_{z_0}^{z_1}dz\int_z^{2T-z}dt\,\partial_t v\left(\eta\partial_t^2u^1-\partial_z\left(\eta\partial_z u^1\right)\right)$$

and repeated use of the boundary value problem (2.1) and the adjoint problem (2.6), yields the identity

$$\left\langle\phi,\ D\tau_0(\sigma,h,f,\eta^0)\cdot(\sigma^1,h^1)\right\rangle=\eta(z_1)\int_{z_1}^{2T-z_1}dt\,\partial_t v(z_1,t)\ h^1(t)$$
$$+\int_{z_0}^{z_1}dz\,\eta(z)\left\{\partial_z\bar{v}\partial_z\bar{u}^1(z)+\sigma^1\left(\int_z^{2T-z}dt\,\partial_t v(z,t)\partial_z u(z,t)\right)\right\}.$$

Now

$$\int_{z_0}^{z_1} dz\,\eta\partial_z\bar{v}\partial_z\bar{u}^1 = -\frac{1}{2}\int_{z_0}^{z_1} dz\,\eta\partial_z\bar{v}\partial_z\left(\eta^{-1/2}\int_{z_0}^{z}\sigma^1\right)$$

$$= -\frac{1}{2}\int_{z_0}^{z_1} dz\,\eta^{1/2}\partial_z\bar{v}\left(\sigma^1 - \frac{1}{2}\sigma\int_{z_0}^{z}\sigma^1\right)$$

$$= -\frac{1}{2}\int_{z_0}^{z_1} dz\,\sigma^1\left\{\eta^{1/2}\partial_z\bar{v} + \frac{1}{2}\int_{z}^{z_1}\sigma\eta^{1/2}\partial_z\bar{v}\right\}$$

whence follows the identity, valid for smooth ϕ with $\phi\,(2T-z_0)=0$,

$$D_{\tau_0}^*(\sigma,h,f,\eta_0)\phi = \left(-\frac{1}{2}\left(\eta^{1/2}\partial_z\bar{v} + \frac{1}{2}\int^{z_1}\sigma\eta^{1/2}\partial_z\bar{v}\right)\right.$$

$$\left. + \eta\int_{z}^{2T-z} dt\,\partial_t v\partial_z u,\ \eta(z_1)\partial_t v(z_1,\cdot)\right).$$

PROPOSITION 2.5. *For C depending on σ, h, f, η^0, T, and z_1-z_0, and $\phi\in C^\infty[z_0,2T-z_0]$ with $\phi(2T-z_0)=0$,*

$$\|\phi\| \leqq C\|D\tau_0^*(\sigma,h,f,\eta_0)\cdot\phi\|.$$

Proof. This argument is very similar to the proof of Lemmas 1.1, 1.3. Set

$$Q(z) = \frac{1}{2}\int_{z}^{2T-z} dt\left\{(\partial_t v)^2 + (\partial_z v)^2\right\}(z,t).$$

From the identity (1.4) we obtain

$$Q(z) = Q(z_1) + \frac{1}{2}\int_{z}^{z_1}|\partial_z\bar{v}|^2 + \int_{z}^{z_1} d\zeta\,\sigma(\zeta)\int_{\zeta}^{2T-\zeta}|\partial_z v(\zeta,t)|^2. \tag{2.7}$$

From (2.6c)

$$Q(z_1) = \int_{z_1}^{2T-z_1}|\partial_t v|^2(z_1,\cdot).$$

On the other hand, as we have assumed σ to be smooth, the operator

$$\psi \to \psi + \frac{1}{2}\eta^{-1/2}\int^{z_1}\sigma\eta^{1/2}\psi$$

is a Volterra operator of the second kind on $L^2[z_0,z_1]$ with continuous kernel, hence invertible. In particular for suitable $C>0$,

$$\|\partial_z\bar{v}\|_{L^2[z,z_1]} \leqq C\left\|-\frac{1}{2}\left(\eta^{1/2}\partial_z\bar{v} + \frac{1}{2}\int^{z_1}\sigma\eta^{1/2}\partial_z\bar{v}\right)\right\|_{L^2[z,z_1]}$$

for $z_0\leqq z\leqq z_1$.

On the other hand,

$$\left(-\frac{1}{2}\left(\eta^{1/2}\partial_z\bar{v} + \frac{1}{2}\int^{z_1}\sigma\eta^{1/2}\partial_z\bar{v}\right),\ \eta(z_1)\partial_t v(z_1,\cdot)\right)$$

$$= D\tau^*(\sigma,h,f,\eta_0)\cdot\phi - \left(\eta\int_{z}^{2T-z} dt\,\partial_t v\partial_z u,\ 0\right)$$

and

$$\int_z^{2T-z} dt\,\partial_t v\,\partial_z u \leqq CQ^{1/2}(z)$$

from Cauchy–Schwarz and Proposition 2.1. Thus

$$\begin{aligned} Q(z_1)+\frac{1}{2}\int_z^{z_1}|\partial_z\bar{v}|^2 &\leqq C\left\{\eta(z_1)^2\int_{z_1}^{2T-z_1}|\partial_t v(z_1,\cdot)|^2\right. \\ &\qquad \left. +\frac{1}{4}\int_z^{z_1}\left(\eta^{1/2}\partial_z\bar{v}+\frac{1}{2}\int^{z_1}\sigma\eta^{1/2}\partial_z\bar{v}\right)^2\right\} \\ &\leqq C\left\{\left\|D\tau^*(\sigma,h,f,\eta_0)\cdot\phi\right\|+\int_z^{z_1}d\zeta\left(\eta(\zeta)\int_\zeta^{2T-\zeta}dt\,\partial_t v\,\partial_z u\right)^2\right\} \\ &\leqq C_1\left\{\left\|D\tau^*(\sigma,h,f,\eta_0)\cdot\phi\right\|^2+C_2\int_z^{z_1}Q\right\}. \end{aligned}$$

Thus we can replace the RHS of (2.7) by

$$Q(z)\leqq C_1\left\|D\tau^*(\sigma,h,f,\eta_0)\cdot\phi\right\|^2+2\int_z^{z_1}(2|\sigma|+C_2)Q$$

whence from Gronwall's inequality

$$\|\phi\|\leqq 2Q(z_0)\leqq C\left\|D\tau^*(\sigma,h,f,\eta_0)\cdot\phi\right\|^2. \qquad \text{Q.E.D.}$$

COROLLARY 2.6. *For smooth* $(\sigma,h,f,\eta_0)\in K_D^\infty$, $D\tau(\sigma,h,f,\eta_0)$ *is a linear isomorphism:* $H_D\to H_T$. *Further, the bounds for* $\|D\tau\|$, $\|D_\tau^{-1}\|$ *depend only on* $\|\sigma\|$, $\|h\|$, $\|f\|$, η_0, T, *and* z_1-z_0.

Proof. The previous proposition shows that the adjoint $D\tau^*$ is injective, since the $\phi\in C^\infty[z_0,2T-z_0]$ with $\phi(2T-z_0)=0$ are dense in H_T. Therefore, the range of $D\tau$ is dense. We have already noted that it is closed: so $D\tau$ is surjective. Proposition 2.4 gave the required bounds. Q.E.D.

THEOREM 2.7. $D\tau$ *is a locally, Lipschitz-continuous map on* K_D^∞ *with values in* $L(H_D,H_T)$, *whence* $D\tau$ *extends to a locally Lipschitz map*

$$K_D\times H_D\to H_T.$$

Proof. Select (σ,h,f_0,η_0) and $(\tilde{\sigma},\tilde{h},\tilde{f}_0,\tilde{\eta}_0)$ in K_D^∞, and let u, $\tilde{u}$ be the corresponding solutions of (2.1). For (smooth) (σ^1,h^1) let u^1 and $\tilde{u}^1$ be the corresponding solutions of (2.4). The difference $v=\tilde{u}^1-u^1$ solves

$$(2.8a)\qquad \left(\partial_t^2-\partial_z^2-\sigma\partial_z\right)v=\sigma^1(\partial_z\tilde{u}-\partial_z u)+(\tilde{\sigma}-\sigma)\partial_z\tilde{u}^1 \quad \text{in } \Omega,$$

$$(2.8b)\qquad \partial_z v(z_0,\cdot)=0,$$

$$(2.8c)\qquad (\partial_z+\partial_t)v(z_0,\cdot)=0,$$

$$(2.8d)\qquad \bar{v}(z)=-\frac{1}{2}(\tilde{\eta}^{-1/2}-\eta^{1/2})(z)\int_{z_0}^z\sigma^1.$$

Note that

$$\left\|\partial_z\cdot\left(\tilde\eta^{-1/2}-\eta^{-1/2}\right)\int\sigma^1\cdot\right\| = \left\|\left(\partial_z\left(\tilde\eta^{-1/2}-\eta^{-1/2}\right)\right)\int\sigma^1\right\| + \left\|\left(\tilde\eta^{-1/2}-\eta^{-1/2}\right)\sigma^1\right\|$$

$$\leqq \left\|\partial_z\left(\tilde\eta^{-1/2}-\eta^{-1/2}\right)\right\|\left\|\int\sigma^1\right\|_\infty + \left\|\tilde\eta^{-1/2}-\eta^{-1/2}\right\|_\infty\left\|\sigma^1\right\|$$

$$\leqq C\|\tilde\sigma-\sigma\|\|\sigma^1\|.$$

Now Proposition 2.2 states that, for suitable C,

$$\int_z^{2T-z} dt\cdot\partial_z(\tilde u-u)(z,t)\cdot^2 \leqq C\left\{\|\tilde\sigma-\sigma\|^2+\|\tilde f-f\|^2+\|\tilde h-h\|^2+|\tilde\eta_0-\eta_0|^2\right\}.$$

Implicit in the proof of the Theorem 2.4 is the estimate

$$\int_z^{2T-z} dt\left|\partial_z\tilde u^1\right|^2(z,t)\leqq C\left\{\|\sigma^1\|^2+\|h^1\|^2\right\},$$

where C depends on $\|\tilde f_0\|$, $\|\tilde\sigma\|$, $\tilde\eta_0$, and z_1-z_0. Clearly Lemma 1.4 extends to the case in which the RHS of the wave equation is given by a sum:

$$\sum_{i=1}^N \omega_i w_i.$$

Taking $N=2$, $\omega_1=\sigma^1$, $w_1=\partial_z(\tilde u-u)$, $\omega_2=\tilde\sigma-\sigma$, $w_2=\partial_z\tilde u^1$, gives the estimate

$$\|\partial_t v(z_0,\cdot)\|\leqq C\left(\|\sigma^1\|^2+\|h^1\|^2\right)^{1/2} \cdot\left(\|\sigma-\tilde\sigma\|^2+\|h-\tilde h\|^2+\|f_0-\tilde f_0\|^2+|\eta_0-\tilde\eta_0|^2\right)^{1/2} \tag{2.9}$$

where C is bounded by a smooth function of $\|\sigma\|$, $\|\tilde\sigma\|$, $\|f_0\|$, $\|\tilde f_0\|$, η_0, $\tilde\eta_0$, and z_1-z_0. For any smooth σ, $\tilde\sigma$, etc. belonging to a bounded set in K_D, the inequality (2.9) is therefore an explicit local Lipschitz estimate. Q.E.D.

THEOREM 2.8. *For each $(\sigma,h,f,\eta_0)\in K_D$ there exist $k,r>0$ so that for $(\sigma^1,h^1)\in K_D$ with $\|\sigma^1\|^2+\|h^1\|^2=1$, and all $\varepsilon<r$,*

$$\left\|D\tau_0(\sigma,h,f_0,\eta_0)\cdot(\sigma^1,h^1)-\frac{1}{\varepsilon}\left[\tau_0\left(\sigma+\varepsilon\sigma^1,h+\varepsilon h^1,f,\eta_0\right)-\tau_0(\sigma,h,f,\eta_0)\right]\right\|<k\varepsilon.$$

Proof. Let $\tilde\sigma=\sigma+\varepsilon\sigma^1$, $\tilde h=h+\varepsilon h^1$,

$$\tilde\eta(z)=\eta_0\exp\int_{z_0}^z\tilde\sigma,\qquad \eta(z)=\eta_0\exp\int_{z_0}^z\sigma.$$

Let u solve (2.1), $\tilde u$ solve (2.1) with η, h replaced by $\tilde\eta$, $\tilde h$ and u^1 solves (2.4). Then the quantity we need to estimate is

$$\|\partial_t v(z_0,\cdot)\|$$

where

$$v=u^1-\varepsilon^{-1}(\tilde u-u)$$

solves the boundary value problem

$$(2.10a) \qquad \left(\partial_t^2 - \partial_z^2 - \sigma\partial_z\right)v = \sigma^1\partial_z(u-\tilde{u}),$$

$$(2.10b) \qquad \partial_z v(z_0,\cdot) = 0,$$

$$(2.10c) \qquad (\partial_t + \partial_z)v(z_1,\cdot) = 0,$$

$$(2.10d) \qquad \bar{v} = -\frac{1}{2}\eta^{-1/2}\sigma^1 - \frac{1}{\varepsilon}\left(\tilde{\eta}^{-1/2} - \eta^{-1/2}\right).$$

Note that

$$(2.11) \qquad \tilde{\eta}^{-1/2}(z) - \eta^{-1/2}(z) = \eta^{-1/2}(z)\left(\exp - \frac{\varepsilon}{2}\int_{z_0}^{z}\sigma^1 - 1\right)$$

$$= -\frac{\varepsilon}{2}\eta^{-1/2}(z)\int_{z_0}^{z}\sigma^1 + r(\varepsilon,z).$$

Since $\|\sigma^1\|_{L^2} \leqq 1$, $(1/\varepsilon^2)r(\varepsilon,z)$ is uniformly bounded as $\varepsilon \to 0$ by a function of η_0, $\|\sigma\|$, and $(z_1 - z_0)$. Also,

$$\frac{d}{dz}r(\varepsilon,z) = \left(\frac{d}{dz}\eta^{-1/2}(z)\right)\left\{\exp\left[\left(-\frac{\varepsilon}{2}\int_{z_0}^{z}\sigma^1\right) - 1\right] + \frac{\varepsilon}{2}\int_{z_0}^{z}\sigma^1\right\}$$

$$+ \eta^{-1/2}(z)\left\{-\frac{\varepsilon}{2}\sigma^1(z)\left[\exp\left(-\frac{\varepsilon}{2}\int_{z_0}^{z}\sigma^1\right) - 1\right]\right.$$

$$\left. - \frac{1}{2}\sigma(z)r(\varepsilon,z) - \frac{\varepsilon}{2}\sigma^1(z)\left(\tilde{\eta}^{-1/2}(z) - \eta^{-1/2}(z)\right)\right\}.$$

So (from (2.11))

$$(2.12) \qquad \left\|\frac{dr}{dz}\right\| = O(\varepsilon^2)$$

where the constant implicit in the Landau symbol depends on η_0, $\|\sigma\|$, and $z_1 - z_0$. For convenience, we shall make this dependence a convention for our use of the Landau symbol for the rest of the proof.

It follows from (2.10d), (2.11), and (2.12) that

$$\left\|\frac{d\bar{v}}{dz}\right\| = O(\varepsilon).$$

Note that the RHS of (2.10a) is of the form $\sigma^1 w$, with $\int|\sigma^1|^2 \leqq 1$ and

$$w = \partial_z(u - \tilde{u})$$

satisfying

$$\int_z^{2T-z} dt\,(w(z,t))^2 = O(\varepsilon^2)$$

according to Proposition 2.2. We are once again in position to apply Lemma 1.4. We conclude that

$$\|\partial_t v(z_0,\cdot)\| = O(\varepsilon). \qquad \text{Q.E.D.}$$

Remark. This last estimate is more than sufficient to establish that τ_0 is differentiable as a function of its first two arguments. In fact, similar estimates show that τ_0 is of class C^2 on K_D, i.e. as a function of all of its arguments. Conceivably, τ_0 is actually smooth.

PROPOSITION 2.9. *Let* $r>0$, $(\sigma^0,h^0,f^0,\eta_0^0)\in K_D$. *Set*

$$B_r=\left\{(\sigma,h)\in H_D:\|\sigma-\sigma^0\|^2+\|h-h^0\|^2\leqq r^2\right\}.$$

Then there exists $C=C(\sigma^0,h^0,f^0,\eta_0^0,r,z_1-z_0)$ *so that for any* $(\sigma,h)\in B_r$,

$$\|\sigma\|^2+\|h\|^2\leqq C\left\{\|\tau(\sigma,h,f^0,\eta_0^0)\|^2+\|f^0\|^2\right\}.$$

Proof. It follows immediately from Lemma 1.2 that

$$\|h\|^2+\frac{1}{2}\int_{z_0}^{z_1}(\bar{u}')^2\leqq C\left\{\|\tau(\sigma,h,f^0,\eta_0^0)\|^2+\|f^0\|^2\right\}$$

where C depends on $\|\sigma\|\leqq r+\|\sigma^0\|$, η_0^0, and z_1-z_0. On the other hand

$$\bar{u}'=(\eta^{-1/2})'=\left[(\eta_0^0)^{-1/2}\exp\left(-\frac{1}{2}\int\sigma\right)\right]'=-\frac{1}{2}\eta^{-1/2}\sigma$$

and

$$\begin{aligned}\eta^{-1/2}&\leqq(\eta_0^0)^{-1/2}\exp\left(-\frac{1}{2}\int_{z_0}^{z_1}|\sigma|\right)\\&\leqq(\eta_0^0)^{-1/2}\exp\left(-\frac{1}{2}(z_1-z_0)^{1/2}\|\sigma\|\right)\\&\leqq(\eta_0^0)^{-1/2}\exp\left\{-\frac{1}{2}(z_1-z_0)^{1/2}(r+\|\sigma^0\|)\right\}\end{aligned}$$

whence the conclusion follows. Q.E.D.

THEOREM 2.10. *Let* $(\sigma^0,h^0,f^0,\eta_0^0)\in K_D$. *Then there exist constants*

$$\begin{aligned}r&=r(\|\sigma^0\|,\|h^0\|,\|f^0\|,\eta_0,z_1-z_0),\\C&=C(\|\sigma^0\|,\|h^0\|,\|f^0\|,\eta_0^0,r,z_1-z_0)\end{aligned}$$

so that if $r<R$, *then*

$$|\eta_0-\eta_0^0|^2+\|f-f^0\|^2+\|g-\tau_0(\sigma^0,h^0,f^0,\eta_0^0)\|^2<r^2$$

implies that there exists a unique solution (σ,h) *of the equation*

$$\tau_0(\sigma,h,f,\eta_0)=g\in H_T. \tag{2.13}$$

Moreover,

$$\|\sigma-\sigma^0\|^2+\|h-h^0\|^2\leqq C\left\{(\eta_0-\eta_0^0)^2+\|f-f^0\|^2+\|g-\tau_0(\sigma^0,h^0,f^0,\eta^0)\|^2\right\}.$$

Proof. From Corollary 2.6, $D\tau$ is an isomorphism for smooth data. From Theorem 2.7, the lower bounds of Corollary 2.6 and Proposition 2.5 for $D\tau$ and $D\tau^*$ respectively survive the extension to $K_D\times H_D$, so $D\tau$ is a linear isomorphism for all

$(\sigma,h,f,\eta_0)\in K_D$. Now the implicit mapping theorem guarantees the local existence and continuous dependence on g, f, η_0 of solutions of the functional equation (2.13).

The uniqueness of global solutions of (2.13) follows from (a slight extension of) the results of [21]. We prefer to give an alternative proof more closely related to the arguments which we have extended to several-dimensional problems—see [22].

Suppose $(\tilde\sigma,\tilde h)\in H_D$ with $\tau(\tilde\sigma,\tilde h f,\eta_0)=\tau(\sigma,h,f,\eta_0)$. Denoting by $\tilde u$ and the relevant solutions of (2.1), we see that the difference $v=\tilde u-u$ solves

$$(\partial_t^2-\partial_z^2-\sigma\partial_z)v=(\tilde\sigma-\sigma)\partial_z\tilde u, \tag{2.14a}$$

$$\partial_z v(z_0,\cdot)=0, \tag{2.14b}$$

$$(\partial_z+\partial_t)v(z_1,\cdot)=\tilde h-h, \tag{2.14c}$$

$$\bar v(z)=\tilde\eta^{-1/2}(z)-\eta^{-1/2}(z). \tag{2.14d}$$

Write, $v=v_{\mathrm I}+v_{\mathrm{II}}$, where $v_{\mathrm I}$ solves

$$(\partial_t^2-\partial_z^2-\sigma\partial_z)v_{\mathrm I}=0$$

and (2.14b, c, d) while v_{II} solves (2.14a, b, c) and

$$\bar v_{\mathrm{II}}\equiv 0.$$

Consider for the moment (2.14) restricted to the triangle $\{(z,t): z_0\leqq z\leqq z_2,\ z\leqq t\leqq 2z_2-z\}$, where $z_2\in(z_0,z_1)$, for which the right-hand boundary condition (2.14c) plays no role. Define

$$Q_{\mathrm I}^0(z)=\int_z^{2z_2-z}dt\left\{(\partial_t v_{\mathrm I})^2+(\partial_z v_{\mathrm I})^2\right\}(z,t)$$

and similarly for Q_{II}^0. Since the Cauchy data for v vanish on $\{z=z_0\}$ by hypothesis, we have

$$Q_{\mathrm I}^0(z_0)=Q_{\mathrm{II}}^0(z_0).$$

According to Lemma 1.1,

$$\int_z^{2z_2-z}dt(\partial_z\tilde u)^2(z,t)\leqq C(\eta_0,z_2-z_0,\|\tilde\sigma\|)\|\tilde\sigma\|^2$$

whence from Lemma 1.2

$$Q_{\mathrm{II}}^0(z_0)\leqq C(\eta_0,z_2-z_0,\|\sigma\|,\|\tilde\sigma\|)\|\sigma-\tilde\sigma\|^2\|\tilde\sigma\|^2(z_2-z_0).$$

From Lemma 1.5

$$\int_{z_0}^{z_2}dz\left(\partial_2(\eta^{-1/2}-\tilde\eta^{-1/2})\right)^2\leqq C(\eta_0,z_2-z_0,\|\sigma\|)Q_{\mathrm I}^0(z_0)$$

so

$$|\eta^{-1/2}-\tilde\eta^{-1/2}|\leqq C(\eta_0,z_2-z_0,\|\sigma\|)Q_{\mathrm I}^0(z_0)^{1/2}.$$

Also

$$\sigma-\tilde\sigma=-2\eta^{1/2}\partial_z(\eta^{-1/2}-\tilde\eta^{-1/2})-(1-\eta^{1/2}\tilde\eta^{-1/2})\tilde\sigma$$

whence

$$\begin{aligned}\|\sigma-\tilde\sigma\|^2&\leqq C(\eta_0,z_2-z_0,\|\sigma\|,\|\tilde\sigma\|)Q_{\mathrm I}^0(z_0)\\&\leqq C(\eta_0,z_2-z_0,\|\sigma\|,\|\tilde\sigma\|)\|\sigma-\tilde\sigma\|^2(z_2-z_0).\end{aligned}$$

Now C is continuous in $z_2 - z_0 \to 0$, so for $z_2 - z_0$ small enough, we obtain $\sigma = \tilde{\sigma}$ $p.p.$ on $[z_0, z_2]$. Consequently the equation (2.14a) becomes homogeneous in the slab $\{z_0 \leqq z < z_2\}$, and from Lemma 1.5 we conclude that $v \equiv 0$ in this slab, in particular that v and $\partial_z v$ vanish for $z - z_2$, $z_2 \leqq t \leqq 2T - z_2$. Now we can replace z_0 by z_2 and repeat the argument. After finitely many repetitions we obtain $\sigma - \tilde{\sigma}$ on $[z_0, z_1]$, i.e. we have established global uniqueness for (2.13). Q.E.D.

Remark. This "downward continuation" or "layer-stripping" argument of course gives a (direct) uniqueness result for the (global) inverse problem, i.e. the functional equation (0.4) with $f = -\delta$.

As before, with the special choice $z_0 = 0$, $z_1 = T$, $f \equiv 0$, $\eta_0 = 1$, we obtain Theorem 0.3 as a corollary of Theorem 2.10. Since Theorem 0.2 is implied by Theorem 0.3, we have completed the proofs of our results.

REFERENCES

[1] A. BAMBERGER, G. CHAVENT, AND P. LAILLY, *About the stability of the inverse problem in* 1-D *wave equations-applications to the interpretation of seismic profiles*, Appl. Math. Opt., 5 (1979), pp. 1–47.

[2] A. BRUCKSTEIN, B. C. LEVY, AND T. KAILATH, *Differential methods in inverse scattering*, SIAM J. Appl. Math., 45 (1984), pp. 312–335.

[3] K. BUBE, *Convergence of difference methods for one-dimensional inverse problems*, International Geophysics and Remote Sensing Symposium Digest, IEEE, 1983.

[4] K. BUBE AND R. BURRIDGE, *The one-dimensional inverse problem of reflection seismology*, SIAM Rev., 25 (1983), pp. 497–559.

[5] R. BURRIDGE, *The Gel'fand–Levitan, the Marchenko, and the Gopinath–Sondhi integral equations of inverse scattering theory, regarded in the context of inverse impulse-response problems*, Wave Motion, 2 (1980), pp. 305–323.

[6] R. W. CARROLL AND F. SANTOSA, *Stability for the one-dimensional inverse problem via the Gel'fand–Levitan equation*, Applicable Analysis, 13 (1982), pp. 271–277.

[7] J. COHEN AND N. BLEISTEIN, *An inverse method for determining small variations in propagation speed*, SIAM J. Appl. Math., 32 (1977), pp. 784–799.

[8] R. COURANT AND D. HILBERT, *Methods of Mathematical Physics* II, Wiley Interscience, New York, 1962.

[9] K. DRIESSEL AND W. SYMES, *Fast and accurate algorithms for impedance profile inversion*, Amoco Production Co. Technical Report, 1981.

[10] F. G. FRIEDLANDER, *Sound Pulses*, Cambridge Univ. Press, Cambridge, 1958.

[11] M. GERVER, *Inverse problem for the one-dimensional wave equation*, Geophys. J. Roy. Astr. Soc., 21 (1970), pp. 337–357.

[12] P. GOUPILLAUD, *An approach to inverse filtering of near-surface layer effects from seismic records*, Geophysics, 26 (1961), pp. 754–760.

[13] S. H. GRAY, *A second-order procedure for one-dimensional velocity inversion*, SIAM J. Appl. Math., 39 (1980), pp. 456–462.

[14] O. HALD, *The inverse Sturm–Liouville problem for symmetric potentials*, Acta Math. (Sweden), 141 (1979), pp. 264–291.

[15] G. KUNETZ, *Generalisation des operateurs d'antiresonance à une nombre quelconque de reflecteurs*, Geoph. Prosp., 12 (1964), pp. 283–289.

[16] J. RAUCH AND M. TAYLOR, *Exponential decay of solutions to hyperbolic equations in bounded domains*, Indiana J. Math., 24 (1974), pp. 79–86.

[17] F. SANTOSA AND H. SCHWETLICK, *The inversion of acoustical impedance profile by methods of characteristics*, Wave Motion, 4 (1982), pp. 99–110.

[18] F. SANTOSA AND W. SYMES, *Inversion of impedance profile from band limited data*, International Geophysics and Remote Sensing Symposium Digest, IEEE, 1983.

[19] H. SCHWETLICK, *Inverse methods in the reconstruction of acoustical impedance profiles*, Thesis, Dept. of Theoretical and Applied Mechanics, Cornell Univ., Ithaca, NY, 1982.

[20] M. M. SONDHI AND J. R. RESNICK, *The inverse problem for the vocal tract: numerical methods, acoustical experiments, and speech synthesis*, J. Acoust. Soc. Amer., 73 (1983), pp. 985–1002.

[21] W. SYMES, *Impedance profile inversion via the first transport equation*, J. Math. Anal. Appl., 84 (2) (1983), pp. 435–453.

[22] _____, *Linearization stability for an inverse problem in several-dimensional wave propagation*, this Journal, pp. 132–151.

[23] _____, *Inverse Problems in Several-Dimensional Wave Propagation*, International Geophysics and Remote Sensing Symposium Digest, IEEE, 1983.

[24] _____, *Stable solution of the inverse reflection problem for a smoothly stratified elastic medium*, this Journal, 12 (1981), pp. 421–453.

[25] _____, *Continuation for solutions of wave equations: regularization of the time-like Cauchy problem*, preprint, 1984.

[26] J. WARE AND K. AKI, *Continuous and discrete inverse scattering problems in a stratified elastic medium*, J. Acoust. Soc. Amer., 45 (1969), pp. 911–921.

SIAM J. MATH. ANAL.
Vol. 17, No. 6, November 1986

010

HOMOGENIZATION FOR A VOLTERRA EQUATION*

HEDY ATTOUCH† AND ALAIN DAMLAMIAN‡

Abstract. A model is given for the nonlinear heat equation in a heterogeneous medium with memory. Its homogenization is carried out in two particular cases (including the linear one).

Key words. Volterra equation, homogenization, heat flow, heterogeneous material with memory

AMS(MOS) subject classifications. Primary 45D05, 73K20, 80A20, 45G10, 47H05

1. Introduction. In a heterogeneous medium with memory, a model for the heat equation (see Nohel [1]) is

$$\frac{\partial \xi}{\partial t} + \operatorname{div}_x Q = h_1, \tag{1.1}$$

where h_1 is a given diffused source term, ξ is the internal energy, and Q is the heat flux. The latter are assumed to be functionals of the temperature distribution u with "memory":

$$\xi(t,x) = b_0(x)u(t,s) + \int_{-\infty}^{t} \beta(x,t-s)u(s,x)\,ds, \tag{1.2}$$

$$Q(t,x) = -c_0(x)\sigma(x,\nabla u(t,x)) + \int_{-\infty}^{t} \gamma(x,t-s)\sigma(x,\nabla u(s,x))\,ds. \tag{1.3}$$

Equation (1.1) is considered on the product $\Omega \times (-\infty, T)$, where Ω is a bounded regular domain in R^3 (or R^N); the function σ: $\Omega \times R^N \to R^N$, $(x,r) \to \sigma(x,r)$ represents a nonlinear flux law. Its dependence upon x specifies the heterogeneity of the medium. Similarly $b_0(x)$, $\beta(x,T)$, $c_0(x)$, $\gamma(x,t)$ characterize the spatial heterogeneity of the other thermodynamical parameters.

To equation (1.1) are added boundary and initial conditions which will be specified later.

The questions considered here are:

—under what suitable set of hypotheses is equation (1.1) well posed (existence and uniqueness);

—under what further conditions can one treat the corresponding homogenization problem; in other words, if all parameters involved $(\sigma, b_0, \beta, c_0, \gamma)$ depend upon another variable ε measuring the "tightness" of the heterogeneity of the medium (typically $b_0^\varepsilon(x) = b_0(x/\varepsilon)$ where $b_0(y)$ is periodic), one can find a limit problem whose solution would be the limit of the solutions u_ε, and whose structure would be similar to (1.1), (1.2), (1.3)? We will give positive answers to both questions in particular cases only. The paper is organized as follows:

* Received by the editors February 9, 1984, and in revised form July 23, 1985. This research was sponsored by the U.S. Army under contract DAAG29-80-C-0041.

† Laboratoire de Mathématiques, Université Paris Sud, 91405 Orsay Cedex, France.

‡ Centre de Mathématiques, Ecole Polytechnique, 91128 Palaiseau Cedex, France.

In the next section (§2), we reformulate the problem in terms of a Volterra equation (i.e. with memory convolution kernels), and we make precise the hypotheses on the nonlinear term σ.

In §3 we then study what we called the splitting case, when c/c_0 is independent of the space variable x. In this situation one can apply the nice method of Crandall and Nohel [1] in order to transform the problem into a Lipschitz perturbation of a variational monotone evolution equation for which sophisticated, albeit known, techniques apply to yield a positive answer to the aforementioned questions. When dealing with the homogenization, we of course make use of various results from that theory, some of which are quite involved (see the quoted references).

In the general situation (§4) we obtain existence and uniqueness via an argument of local monotonicity combined with global estimates, both of which require some extra conditions on the various physical parameters. In §5 we give a first result for homogenization in the same framework; it is however restricted to the linear case and this for the obvious reason that we make use of the Laplace transform together with homogenization theory for complex-valued elliptic problems in order to specify the limit problem. To our knowledge, there is no result pertaining to the homogenization of the nonlinear case; any progress in that direction would be very welcome indeed.

A first approach to this type of problem appeared in Raynal [1].

2. Reformulation of the problem as a Volterra equation. Let $*$ denote the usual convolution with respect to t on $[0, +\infty[$.

As "initial" condition, we assume the history of the medium to be known for t negative. One can then rewrite (1.1), (1.2), (1.3) as:

$$\frac{\partial}{\partial t}[b_0 u + \beta * u] + \operatorname{div}_x(-c_0\sigma + \gamma * \sigma) = h \tag{2.1}$$

where σ stands for $\sigma(\cdot, \nabla u(\cdot,\cdot))$ and the right-hand side h includes the history of the system up to time zero.

It is customary to define

$$c(x,t) = c_0(x) - \int_0^t \gamma(x,s)\,ds \tag{2.2}$$

and to assume that $c(x,t)$ and $b_0(x)$ are strictly positive valued (a physical condition). With these notations (2.1) can be written as

$$b_0 u' - \operatorname{div}_x\{(c * \sigma)'\} = h - \beta_0 u - \beta' * u, \tag{2.3}$$

where "$'$" indicates a time derivative.

The initial condition becomes a Cauchy data at $t=0$:

$$u(0) = u_0 \quad \text{in } L^2(\Omega); \tag{2.4}$$

the boundary condition is taken to be compatible with the operator $-\operatorname{div}_x(\sigma(\nabla u))$, for example

$$u = 0 \quad \text{on } \partial\Omega \times (0,T). \tag{2.5}$$

Problem (2.3), (2.4), (2.5), upon integration with respect to time appears as a Volterra type integral equation (cf. 3.1).

We now make precise the type of function σ appearing here: let j: $(x,r)\in\Omega\times R^N \mapsto j(x,r)\in R^+$, be of Caratheodory type, convex and equicoercive in r. Assume further that $j(x,0)\equiv 0$; let σ be its subdifferential with respect to r and put:

$$\phi(u)=\begin{cases}\int_\Omega j(x,\nabla u(x))\,dx & \text{for } u\in W_0^{1,1}(\Omega),\\ +\infty & \text{otherwise;}\end{cases} \tag{2.6}$$

then it is shown in Attouch-Damlamian [1] that ϕ is lower semi-continuous (l.s.c.) convex on $L^2(\Omega)$; moreover, denoting by $\partial\phi$ the subdifferential of ϕ on $L^2(\Omega)$ one has:

$$v\in\partial\phi(u)\Rightarrow\begin{cases}u\in W_0^{1,1}(\Omega)\cap L^2(\Omega),\\ v=-\operatorname{div}h \text{ for some } h \text{ such that } h(x)\in\sigma(x,\nabla u(x)) \text{ a.e.}\end{cases} \tag{2.7}$$

This is the sense in which σ is used in (2.3) and it also gives a meaning to the Dirichlet condition (2.5) (when j is even with respect to r, (2.7) is actually an equivalence).

3. The splitting case. By this, we mean that c/c_0 is independent of x. Equivalently, by an obvious change of notation (for σ), c and γ can be taken independent of x (c_0 is taken to be 1).

3.1. Existence and uniqueness. Integration of (2.3) with respect to time from 0 to t yields:

$$b_0u-\operatorname{div}_x(c*\sigma)=b_0u_0-\beta*u+H \tag{3.1}$$

where H is the integral of h.

PROPOSITION 3.2. *Assume that b_0 and b_0^{-1} are in $L^\infty(\Omega)^+$, that β is in $L^\infty(\Omega;BV(0,T))$, and that γ is in $BV(0,T)$. Then equation* (3.1) *has a unique solution u in $C([0,T];L^2(\Omega))\cap L^1(0,T,W_0^{1,1}(\Omega))$. Furthermore, du/dt is in $L^2(\Omega\times(0,T))$.*

Proof. We follow here the ideas of Crandall–Nohel [1]. Let e be the resolvant kernel of γ, i.e., the solution of

$$e-\gamma-\gamma*e=0. \tag{3.3}$$

Making use of standard results for convolution equations, one obtains that e belongs to $BV(0,T)$ as soon as γ does so. Using (3.3), (2.3) becomes

$$b_0\frac{du}{dt}-\operatorname{div}_x\sigma(x,\nabla u)=G(u),\qquad u(0)=u_0, \tag{3.4}$$

where

$$G(u)=h+h*e-(\beta_0+b_0e_0)u+b_0u_0e-u*(\beta_0e+b_0e'+\beta'+e*\beta'). \tag{3.5}$$

Above $\beta_0=\beta(0)$, $e_0=e(0)$ $(=\gamma(0))$ and β', e', are the measure derivatives of β and e.

It is easy to check that G is Lipschitz continuous from $L^1(0,t;L^2(\Omega))$ into itself for each $t>0$. In order to apply a fixed point theorem (as in Crandall–Nohel [1]), we first consider the following problem for w in $C([0,T];L^2(\Omega))$:

$$b_0\frac{du}{dt}-\operatorname{div}_x\sigma(x,\partial u)=w,\qquad u(0)=u_0. \tag{3.6}$$

Now, the operator $u \mapsto -(1/b_0)\mathrm{div}_x \sigma(x, \nabla u)$ is the subdifferential of ϕ (see (2.6)) on $L^2(\Omega)$ provided the norm on $L^2(\Omega)$ is chosen with the weight function $b_0(x)$ (since b_0 and b_0^{-1} are in $L^\infty(\Omega)^+$, this is an equivalent Hilbert norm on $L^2(\Omega)$). Therefore, (3.6) can be solved with classical estimates which allow to apply the Lipschitz fixed point theorem to solve (3.4), (3.5). □

3.2. Homogenization. We now assume that $b_0^\varepsilon, c_0^\varepsilon, \gamma^\varepsilon, \beta^\varepsilon$ and σ^ε depend upon an extra parameter ε measuring the size of the heterogeneity of the medium. A typical example is the periodic case where $b_0^\varepsilon(u) = \tilde{b}_0(x/\varepsilon)$, $c_0^\varepsilon(x) = \tilde{c}_0(x/\varepsilon)$, etc. $\cdots$ where $\tilde{b}_0(y)$, $\tilde{c}_0(y), \cdots$ are Y-periodic (Y is an N-dimensional parallelepipedon). We make the following hypotheses:

$$b_0^\varepsilon \text{ and } (b_0^\varepsilon)^{-1} \text{ are bounded in } L^\infty(\Omega)^+, \tag{3.7}$$

β^ε is bounded in $L^\infty(\Omega; BV(0,T))$ and γ^ε is bounded in $BV(0,T)$. Applying Proposition 3.2, one gets

PROPOSITION 3.8. *Under hypothesis* (3.7), *there exists a unique solution* u^ε *for problem* $(3.1)_\varepsilon$ *and* u^ε *is bounded in* $C([0,T]; L^2(\Omega))$ *by a constant involving only* $|h|_{L^2(Q)}$, $|\beta_0^\varepsilon|_{L^\infty(\Omega)}$, $|b_0^\varepsilon|_{L^\infty(\Omega)}$, $|(b_0^\varepsilon)^{-1}|_{L^\infty(\Omega)}$, $|\beta^\varepsilon|_{L^\infty(\Omega; BV(0,T))}$ *and* $|\gamma^\varepsilon|_{BV(0,T)}$. □

In order to study the convergence of u^ε when ε goes to zero, we make the following extra hypotheses (3.9)–(3.13):

(3.9) $j^\varepsilon(x,r)$ is coercive in r uniformly with respect to x and ε; more precisely: $\exists K_1 > 0$, $\exists K_2 \in R$ and $p > 2N/(N+2)$ such that $j^\varepsilon(x,r) > K_1|r|^{2N/(N+2)} - K_2$.

(3.10) ϕ^ε (defined as in (2.6)) converge in the sense of Mosco[1] on $L^2(\Omega)$ to a limit denoted ϕ, which is then known to be of the same integral form, associated to a convex function j (cf. Attouch [1]). Finally $\phi(u_0)$ is assumed to be finite.

(3.11) β^ε converges in the weak star topology of $L^\infty(\Omega; BV(0,T))$ to a limit β (consequently β_0^ε converges to β_0 in the weak star topology of $L^\infty(\Omega)$; in the periodic case, β is just the average of β^ε over a period).

(3.12) b_0^ε converges to some b_0 in the weak star topology of $L^\infty(\Omega)$ and e^ε converges to some e in the weak star topology of $BV(0,T)$. Then, e is the resolvant of some γ in $BV(0,T)$ (but γ has no relationship to the weak star limit of γ^ε in $BV(0,T)$).

Consequently, the mapping G^ε (the analogue of G in (3.5)) is bounded so that $G^\varepsilon(u^\varepsilon)$ is bounded in $L^2(0,T; L^2(\Omega))$ and therefore, via the properties of the solutions of the problems $(3.6)_\varepsilon$, one can conclude (making also use of (3.10)) that

(3.13) u^ε is bounded in $L^\infty(0,T; W_0^{1,p}(\Omega))$, du^ε/dt is bounded in $L^2(0,T; L^2(\Omega))$; hence, as a consequence of Aubin's lemma (see Aubin [1]), u^ε is compact in $C([0,T]; L^2(\Omega))$.

We will show now that u^ε has only one possible limit value u when ε goes to zero, which implies that u^ε converges to u. Let therefore ε_n go to zero so that u^{ε_n} converges to some u in $C([0,T]; L^2(\Omega))$:

PROPOSITION 3.14. *Under the above hypotheses* (3.13), $G^{\varepsilon_n}(u^{\varepsilon_n})$ *converges weakly in* $L^2(\Omega \times (0,T))$ *to* $G(u)$ (*as given in* (3.5)), *and* $b_0^{\varepsilon_n}\, du^{\varepsilon_n}/dt$ *converges weakly in* $L^2(\Omega \times (0,T))$ *to* $b_0 du/dt$.

[1] For the definition and properties of this convergence, see Mosco [1] or Attouch [2].

Proof. We write ε instead of ε_n for simplicity. $b_0^\varepsilon du^\varepsilon/dt$ is bounded in $L^2(Q)$ ($Q=\Omega\times(0,T)$) and for $\varphi(x,T)$ in $D(Q)$ one has:

$$\int_Q b_0^\varepsilon \frac{du^\varepsilon}{dt}\varphi = -\int_Q b_0^\varepsilon \frac{\partial\varphi}{\partial t}u^\varepsilon \to -\int_Q b_0 \frac{\partial\varphi}{\partial t}u$$

which proves the second claim. For $G^\varepsilon(u^\varepsilon)$, each term can be treated independently; let us show convergence for the worst case: $w^\varepsilon=\beta^{\varepsilon\prime} * u^\varepsilon * e^\varepsilon$. First

$$w^\varepsilon(x,t)=\int_0^t e^\varepsilon(t-s)\int_0^s u^\varepsilon(x,s-\sigma)\,d\beta^\varepsilon(x,\sigma)\,ds$$

so that

$$|w^\varepsilon(x,\cdot)|_{L^\infty(0,T)}\leqq|e^\varepsilon|_{L^1(0,T)}|u^\varepsilon(x,\cdot)|_{L^\infty(0,T)}|\beta^\varepsilon(x,\cdot)|_{BV(0,T)},$$

and

$$|w^\varepsilon|_{L^2(\Omega;\,L^\infty(0,T))}<|e^\varepsilon|_{L^1(0,T)}|u^\varepsilon|_{L^2(\Omega;\,L^\infty(0,T))}|\beta^\varepsilon|_{L^\infty(\Omega;\,BV(0,T))}.$$

Consequently, w^ε is bounded in $L^2(\Omega;L^\infty(0,T))$ because of u^ε being bounded in $L^2(\Omega;H^1(0,T))$ is bounded in $L^2(\Omega;C([0,T]))$. By a similar argument, one checks that for almost every x in Ω, $u^\varepsilon(x,t)\to u(x,t)$ in $C([0,T])$ which will be enough to show the convergence of w^ε to the proper w in $\mathscr{D}'(Q)$ as follows: for φ in $\mathscr{D}(Q)$,

$$\langle w^\varepsilon,\varphi\rangle=-\int_\Omega\iiint\varphi'(x,t+s+\sigma)e^\varepsilon(t)u^\varepsilon(x,s)\beta^\varepsilon(x,\sigma)\,ds\,ds\,d\sigma\,dx$$

$$=\iint dt\,ds\,e^\varepsilon(t)\int\int_\Omega d\sigma\,dx\,\varphi'(x,t+s+\sigma)u^\varepsilon(x,s)\beta^\varepsilon(x,\sigma).$$

For every (t,s) the last integral converges to

$$\int d\sigma\int_\Omega\varphi'(x,t+s+\sigma)u(x,s)\beta(x,\sigma)\,dx$$

because β^ε converges to β $*$-weakly in $L^\infty(Q)$. Furthermore, Lebesgue's dominated convergence theorem applies to the (t,s) integral since the integrand is actually bounded by a constant, namely

$$|\varphi'|_\infty\cdot\sup|u^\varepsilon|_{L^\infty(0,T;\,L^2(\Omega))}\cdot\sup|\beta^\varepsilon|_{L^\infty(Q)},$$

this last factor being bounded above by $\sup|\beta^\varepsilon|_{L^\infty(\Omega,\,BV(0,T))}$ which is finite by hypothesis (3.7). □

Making now use of the convergence in the sense of Mosco of ϕ^ε to ϕ which implies a demi-closedness property (cf. Attouch [2]) one passes to the limit in

$$-\operatorname{div}\sigma^\varepsilon(x,\nabla x^\varepsilon)=G^\varepsilon(u^\varepsilon)-b_0^\varepsilon\frac{du^\varepsilon}{dt}$$

to conclude that u is a solution of

$$-\operatorname{div}\sigma(x,\nabla u)=G(u)-b_0\frac{du}{dt},\qquad u(0)=u_0. \tag{3.15}$$

Because G is as in (3.5), one concludes to the uniqueness of the solution for (3.15), hence the conclusion:

THEOREM 3.16. *Under the hypotheses* (3.7), (3.9), (3.10), (3.11), (3.12) (*and with the notation therein*), *the solution* u^ε *of problem* $(3.4)^\varepsilon$, $(3.5)^\varepsilon$ *converges uniformly in* $C([0,T];L^2(\Omega))$ *to the solution* u *of the analogous problem whose thermodynamical parameters are obtained as follows*:

β *and* b_0 *are the weak-star limits of* β^ε *and* b_0^ε *respectively*;

σ *is the elliptic homogenization of* σ^ε (*equivalent to the Mosco convergence of* ϕ^ε *to* ϕ *cf.* (3.10));

γ *is the resolvant of the weak-limit of the resolvant of* γ^ε (*these two operations do not commute*!).

4. Existence and uniqueness in the general case. We start with (2.3) again:

$$b_0u'-\operatorname{div}\big((c*\sigma)'\big)=g(u),\qquad u(0)=u_0, \tag{4.1}$$

where

$$g(u)=h-\beta_0-\beta'*u \tag{4.2}$$

is Lipschitz continuous from $L^1(0,t;L^2(\Omega))$ into itself for every positive t. Therefore, one can solve (4.1) as a Lipschitz perturbation problem (in a fashion similar to that in §3.1) provided one can solve

$$b_0\frac{du}{dt}-\operatorname{div}\big((c*\sigma)'\big)=F,\qquad u(0)=u_0 \tag{4.3}$$

for given F, via a monotonicity argument.

Here, one should notice that the method of Crandall–Nohel does not apply because c depends upon x so that div and convolution with c do not commute.

We make the following assumptions where α and k are given positive numbers:

(4.4) i) $b_0, 1/b_0, c_0, 1/c_0, \beta_0, 1/\beta_0$ are bounded by k in $L^\infty(\Omega)$ and $|u_0|_{L^2(\Omega)}<k$.

ii) $\beta'<0$ and $c'=-\gamma<0$ a.e. in Q; $\beta(T,x)$ and $c(T,x)$ are bounded below away from zero by α; $t\mapsto c(t,x)$ is continuous at $t=0$ with values in $L^\infty(\Omega)$.

iii) β'' and $c''(=-\gamma')$ are nonnegative measures for a.e. x in Ω.

iv) (2.6) holds with the following inequalities:
$\forall r,s$ in R^N,

$$\alpha|r-s|^2\leqq(\sigma(x,r)-\sigma(x,s),r-s),$$
$$|\sigma(x,r)-\sigma(x,s)|\leqq k|r-s|.$$

We start by choosing t small enough as follows:

PROPOSITION 4.5. *Let* $Au=-\operatorname{div}((c*\sigma)')=-\operatorname{div}(c_0\sigma)-\operatorname{div}(c'*\sigma)$. *For* T *small enough*, A *is maximal monotone from* $L^2(0,T;H_0^1(\Omega))=V$ *into* $L^2((0,T);H^{-1})=V'$.

Proof. We estimate

$$
\begin{aligned}
&(Au-AV,u-v)_{V',V}\\
&\quad\geqq\int_Q c_0(\sigma(x,\nabla u(x))-\sigma(x,\nabla v(x)),\nabla u(x)-\nabla v(x))\,dx\,dt\\
&\qquad+\int_Q (c'*(\sigma,x,\nabla u(x))-\sigma(x,\nabla v(x)),\nabla u(x)-\nabla v(x))\,dx\,dt\\
&\quad\geqq\alpha\int_Q c_0|\nabla u-\nabla v|^2\,dx\,dt\\
&\qquad-\int_Q |c'|*|\sigma(\nabla u)-\sigma(\nabla v)||\nabla u-\nabla v|\,dx\,dt\\
&\quad\geqq\alpha\int_Q c_0|\nabla u-\nabla v|^2\,dx\,dt\\
&\qquad-\int_\Omega dx|c'(x)|_{L^1(0,T)}|\sigma(\nabla u)-\sigma(\nabla v)|_{L^2(0,T)}|\nabla u-\nabla v|_{L^2(0,T)}\\
&\quad\geqq\alpha\int_Q c_0|\nabla u-\nabla v|^2\,dx\,dt-k|c'|_{L^\infty(\Omega;L^1(0,T))}\int_Q|\nabla u-\partial v|^2\,dx\,dt.
\end{aligned}
$$

Note now that $|c'(x)|_{L^1(0,T)}=c_0(x)-c(t,x)$ so by (4.4ii) $|c'|_{L^\infty(\omega;L^1(0,T))}$ is arbitrarily small for T arbitrarily small close to zero. For such a T, $(Au-Av,u-v)>\theta|\nabla u-\nabla v|^2_{L^2(Q)}$ for some positive number θ. Hence A is monotone. Since it is also everywhere defined and continuous on V, it is maximal. □

Now, for such a small T, $b_0d/dt+A$ is one-to-one and onto from V to V' (because of a standard nonlinear argument of coerciveness, cf. Brezis [1]). This proves local existence and uniqueness of the solution for problem (4.3), and hence for (4.1), (4.2). In order to prove global existence, we now get two a priori estimates.

PROPOSITION 4.6. *Under the above hypotheses, there is a constant* C_1 (*depending upon* α *and* k) *such that if* u *is a solution of* (4.1), (4.2), *the following holds*:

$$
|u|_{L^\infty(0,T;L^2(\Omega))}\leqq C_1\quad\text{and}\quad |j(\nabla u)|_{L^1(Q)},\ |j^*(\sigma(\nabla u))|_{L^1(Q)}\leqq C_1
$$

(*consequently* $|\nabla u|^2_{L^2(Q)}\leqq kC_1$, $\sigma|(\partial Vu)|^2_{L^2(Q)}\geqq C_1/\alpha$).

Proof. Multiply (4.1), (4.2) by $u(t)$ and integrate by parts to get

$$
\begin{aligned}
&\frac{1}{2}\frac{d}{dt}|b_0^{1/2}u(t)|^2_{L^2(\Omega)}+|\beta_0^{1/2}u(t)|^2_{L^2(\Omega)}\\
&\qquad+\int_\Omega(\beta'*u)(t)u(t)\,dx+\int_\Omega(c_0\sigma+c'*\sigma)(t)\cdot\nabla u(t)\,dx=\int_\Omega h(t)u(t)\,dx.
\end{aligned}
\tag{4.7}
$$

Integrating on $(0,t)$ yields

$$
\begin{aligned}
&\frac{1}{2}|b_0^{1/2}u(t)|^2_{L^2(\Omega)}+\int_0^t\int_\Omega\beta_0u^2\,ds\,dx+\int_0^t\int_\Omega c_0\sigma(\nabla u(s))\nabla u(s)\,ds\,dx\\
&\qquad=\int_0^t\int_\Omega h\cdot u\,ds\,dx+\int_0^t\int_\Omega(-\beta'*u)(s)u(s)\,ds\,dx\\
&\qquad+\int_0^t\int_\Omega(-c'*\sigma)(s)\nabla u(s)\,ds\,dx+\frac{1}{2}|b_0^{1/2}u_0|^2_{L^2(\Omega)}.
\end{aligned}
\tag{4.8}
$$

By (4.4ii), one has

$$(4.9)\quad \int_0^t (-\beta' * u)(s)u(s)\,ds$$
$$\leqq \int_0^t \int_0^s -\beta'(s-\tau)\left(\frac{1}{2}|u(s)|^2+\frac{1}{2}|u(\tau)|^2\right)ds\,\tau$$
$$\leqq \int_0^t \frac{1}{2}|u(s)|^2(\beta_0-\beta(s))\,ds+\frac{1}{2}\int_0^t ds\int_0^s -\beta'(s-\tau)|u(\tau)|^2\,d\tau$$
$$\leqq \frac{1}{2}\int_0^t (\beta_0-\beta(s))|u(s)|^2\,ds+\frac{1}{2}\left(\int_0^t -\beta'(\tau)d\tau\right)\left(\int_0^t |u(\tau)|^2\,d\tau\right)$$
$$\leqq (\beta_0-\beta(t))\int_0^t u^2(s)\,ds.$$

Hence

$$(4.10)\qquad \int_0^t \int_\Omega \beta_0 u^2+(\beta' * u)u\,dx\,dx \geqq \int_\Omega \beta(x,t)\int_0^t |u(x,s)|^2\,ds\,dx.$$

A similar computation, making use of Young's inequality corresponding to j and j^*, gives

$$(4.11)\qquad \int_0^t \int_\Omega \left(c_0\sigma(\nabla u)+c' * \sigma(\nabla u)\right)\cdot\nabla u\,dx\,ds$$
$$\geqq \int_\Omega c(x,t)\int_0^t j(x,\nabla u(x,s))+j^*(x,\sigma(x,\nabla u(x,s)))\,ds\,dx.$$

Now (4.8), (4.10) and (4.11) combined give

$$\frac{1}{2}|b_0^{1/2}u(t)|^2_{L^2(\Omega)}+\int_\Omega \beta(t,x)\int_0^t u^2(s,x)\,ds\,dx$$
$$(4.12)\qquad +\int_\Omega c(t,x)\int_0^t \left(j(s,\nabla u)+j^*(x,\sigma(\nabla u))\right)ds\,dx$$
$$\leqq \int_0^t \int_\Omega hu\,ds\,dx+\frac{1}{2}|b_0^{1/2}u_0|^2_{L^2(\Omega)}.$$

A standard application of Gronwall's inequality finally yields the desired result.

PROPOSITION 4.13. *Assume the above hypotheses and that $\phi(u_0)$ is finite. Then, there is a constant C_2 (depending on c,k and $\phi(u_0)$) such that whenever u is a solution of* (4.1), (4.2) *then*

$$\left|\frac{du}{dt}\right|_{L^2(Q)}\leqq C_2,\qquad |\phi(u)|_{L^\infty(0,T)}\leqq C_2.$$

Proof. Multiply (4.1), (4.2) by du/dt to get

$$(4.14)\quad \int_0^t \int_\Omega b_0\left|\frac{du}{dt}\right|^2 dx\,dx+\int_0^t \int_\Omega \left(\beta_0 u+\beta' * u\right)\frac{du}{dt}(s)\,dx\,ds$$
$$+\int_0^t \int_\Omega \left(c_0\sigma+c' * \sigma\right)\frac{d}{dt}(\nabla u(s))\,dx\,dx=\int_0^t \int_\Omega h\frac{du}{dt}(s)\,dx\,ds.$$

In (4.14) we integrate by parts the third term as follows:

$$\int_0^t (\beta' * u)u' = u(t)\int_0^t u(s)\beta'(t-s)\,ds - \beta'(0)\int_0^t u^2(s)\,ds$$
$$- \int_0^t u(s)\int_0^s u(\sigma)\beta''(s-\sigma)\,d\sigma\,ds.$$

Making use of $\beta'' \geqq 0$, one can evaluate the last term in a way exactly similar to (4.9) above to get

$$\int_0^t (\beta'' u)\frac{du}{dt} \geqq u(t)\int_0^t u(s)\beta'(t-s)\,ds - \beta'(0)\int_0^t u^2(s)\,ds. \tag{4.15}$$

Again, similarly,

$$\int_0^t (c' * \sigma)\frac{d}{dt}(\nabla u)\,ds \geqq \nabla u(t)\cdot(c' * \sigma)(t) - c'(0)\int_0^t \sigma(s)\nabla u(s)\,ds. \tag{4.16}$$

Using (4.15), (4.16) and the following consequence of the definition of subdifferential $\sigma = \partial j$:

$$\sigma \cdot \frac{d}{dt}(\nabla u) = \frac{d}{dt} j(\nabla u), \tag{4.17}$$

(4.14) yields

(4.18)

$$\int_0^t \int_\Omega b_0 \left|\frac{du}{dt}\right|^2 dx\,ds + \frac{1}{2}\left|\beta_0^{1/2} u(t)\right|^2_{L^2(\Omega)} + \int c_0 j(\nabla u(t))\,ds$$
$$\leqq \int_0^t \int_\Omega h\frac{du}{dt}\,dx\,ds + \frac{1}{2}\left|\beta^{1/2} u_0\right|^2_{L^2(\Omega)} + \int_\Omega c_0 j(\nabla u_0)\,dx$$
$$+ \int_\Omega u(x,t)\int_0^t u(x,s)\big(-\beta'(x,t-s)\big)\,dx\,ds$$
$$+ \int_\Omega \nabla u(x,t)\int_0^t \sigma(x,\nabla u(x,s))\big(-c'(x,t-s)\big)\,ds\,dx.$$

In the right-hand side of (4.18) one can use the following bounds:

$$\text{i)} \int_\Omega \int_0^t u(t)u(s)\big(-\beta'(t-s)\big)\,dx\,ds \leqq |u|^2_{L^\infty(0,T;L^2(\Omega))}\cdot|\beta'|_{L^1(0,T;L^\infty(\Omega))}. \tag{4.19}$$

ii) By Young's inequality

$$\int_\Omega \int_0^t \sigma(s)\nabla u(t)\cdot\big(-c'(t-s)\big)\,ds\,dt$$
$$\leqq \int_\Omega j(\nabla u(t))\big(c_0 - c(t)\big) + |c'|_{L^\infty(Q)}\left|j^*(\sigma(\nabla u))\right|_{L^1(Q)}.$$

Now, confronting (4.18) and (4.19) with hypothesis (4.4ii) and Proposition (4.6), one can conclude. □

THEOREM 4.20. *Under hypotheses* (4.4), *problem* (4.1), (4.2) *has a unique solution on* $[0,T]$.

Proof. From Proposition 4.5, there is existence and uniqueness on some interval $[0,t]t>0$. But the very same proposition gives existence and uniqueness locally in time (starting from $\tau>0$, the problem is changed only insofar as G in (4.2) is modified to incorporate the history up to τ; this in no way changes the conclusion because the a priori estimates are global in time). Combining local existence and the a priori estimates (4.6) and (4.13) gives the result in a standard way. □

5. Homogenization (for the linear case). In this paragraph, b_0^ε, β^ε, c^ε and γ^ε will depend upon the parameter ε which will tend to zero; similarly, σ^ε will depend upon ε but will be assumed to be linear with respect to ∇u; hence the notation

$$\sigma^\varepsilon(x,r)=A^\varepsilon(x)r$$

where $A^\varepsilon(x)$ is a measurable function from Ω to a fixed (independent of ε) set of symmetric uniformly positive definite matrices.

We shall assume hypotheses (4.4) to be satisfied uniformly with respect to ε, so that (4.6), (4.13) and (4.20) hold uniformly in ε.

Consequently, the solutions u^ε of equations $(4.1)_\varepsilon$,$(4.2)_\varepsilon$, belong to a compact set of $C([0,T];L^2(\Omega))$ and a bounded set of

$$L^\infty\big(0,T;H_0^1(\Omega)\big)\cap W^{1,2}\big(0,T;L^2(\Omega)\big).$$

The question of homogenization for $(4.1)_\varepsilon$, $(4.2)_\varepsilon$ is: what can be said of the limit points of u^ε as ε goes to zero?

In order to simplify the notations, we shall assume that ε belongs to a sequence (still denoted ε) such that the following holds:

(5.1) u^ε converges to some u in $C([0,T];L^2(\Omega))$, in the weak-star topology of $L^\infty(0,T;H_0^1(\Omega))$ and the weak topology of $W^{1,2}(0,T;L^2(\Omega))$;

(5.2) b_0^ε converges to b_0 in the weak-star topology of $L^\infty(\Omega)$; β^ε converges to β in

$$\sigma\big(W^{1,1}(0,T;L^\infty(\Omega)),W^{-1,\infty}\big(0,T;L^1(\Omega)\big)\big).$$

Following integration on t, the Volterra equation can be written as

$$-\operatorname{div}\omega^\varepsilon=\mathscr{F}^\varepsilon, \tag{5.3}$$

where

$$\mathscr{W}^\varepsilon(x,t)=\int_0^t A^\varepsilon(x)c^\varepsilon(x,t-s)\nabla u^\varepsilon(x,s)\,ds \tag{5.4}$$

and

$$\begin{aligned}F^\varepsilon(x,t)=&-b_0^\varepsilon(x)u^\varepsilon(x,t)-\int_0^t\beta^\varepsilon(x,t-s)u^\varepsilon(x,s)\,ds\\&+\int_0^t h(x,s)\,ds+b_0^\varepsilon(x)u_0(x).\end{aligned} \tag{5.5}$$

Clearly F^ε converges to F in $C([0,T],H^{-1}(\Omega))$ and weakly in $W^{1,2}(0,T;L^2(\Omega))$, for

$$F(x,t)=b_0(x)(u_0(x)-u(x,t))+\int_0^t h(x,s)\,ds-\int_0^t\beta(x,t-s)u(x,s)\,ds. \tag{5.6}$$

On the other hand, $\mathscr{W}^\varepsilon$ is bounded in $W^{1,2}(0,T;L^2(\Omega))$. In order to characterize the possible weak limits $\mathscr{W}$ of $\mathscr{W}^\varepsilon$ (the uniqueness of which will follow, as usual from the unique solvability of the limit equation), we shall assume, after extracting another subsequence, still denoted ε, that $\mathscr{W}^\varepsilon$ converges weakly to some $\mathscr{W}$. So, going to the limit in (5.3) yields:

$$-\operatorname{div}\mathscr{W}=\mathscr{F}. \tag{5.7}$$

The main task is to find the relationship between $\mathscr{W}$ and u, which (5.4) should yield. Here, because the problem is linear, we use the Laplace transform, but to do so we extend the problem to $[0,+\infty)$ in time as follows: for $t>T$, extend β^ε by $\beta^\varepsilon(x,t)\equiv\beta^\varepsilon(x,T)$, h and γ^ε by zero (so $\sigma^\varepsilon(x,t)\equiv c^\varepsilon(x,T)$), and by Theorem 4.20 which applies to any interval $[0,T_1]$, u^ε exists for all $t\geqq 0$, but for $t>T$, the problems become simpler, as seen from (2.1):

$$b_0^\varepsilon\frac{du^\varepsilon}{dt}+\beta^\varepsilon(T)u^\varepsilon-\operatorname{div}_x A\left(c_0\nabla u^\varepsilon-\gamma*\nabla u^\varepsilon\right)=0.$$

A detailed analysis of estimates (4.6), (4.13) shows that in this particular case, $|u^\varepsilon(t)|_{H^1(\Omega)}$ grows at most exponentially in t with a rate uniform in ε. Therefore, all the Laplace transforms considered here will be convergent at least in some complex right half-plane $\operatorname{Re}\lambda>\lambda_0$.

We will denote by

$$\hat{v}(\lambda)=\int_0^{+\infty}e^{-\lambda t}v(t)\,dt \quad \text{for } \operatorname{Re}\lambda>\lambda_0$$

and (5.3), (5.4) yield $(5.8)_\varepsilon$ below since the gradient operator in x commutes with the Laplace transform.

$$\begin{aligned}&\hat{\mathscr{W}}^\varepsilon(x,\lambda)=A^\varepsilon(x)\hat{c}_\varepsilon(x,\lambda)\nabla\left(\hat{u}_\varepsilon(x,\lambda)\right),\\&-\operatorname{div}\hat{\mathscr{W}}^\varepsilon(x,\lambda)=\hat{\mathscr{F}}^\varepsilon(x,\lambda).\end{aligned} \tag{5.8}_\varepsilon$$

For fixed λ, $(5.8)_\varepsilon$ is just the homogenization problem for an elliptic operator with complex coefficients.

Upon inspection of (5.5), one sees that for every $t>0$, $\mathscr{F}^\varepsilon$ converges to $\mathscr{F}$ in $C([0,t];H^{-1}(\Omega))$ but that $\mathscr{F}^\varepsilon$ grows in H^{-1} at the same rate as u^ε does. Consequently, for each λ with $\operatorname{Re}\lambda>\lambda_0$, $\hat{\mathscr{F}}^\varepsilon(\lambda)$ converges to $\hat{\mathscr{F}}(\lambda)$ in $H^{-1}(\Omega;\mathbb{C})$. Similarly $\hat{u}_\varepsilon(\lambda)$ converges to $\hat{u}(\lambda)$ weakly in $H_0^1(\Omega;\mathbb{C})$. The sesquilinear form

$$a_\varepsilon(\lambda;u,v)=\int_\Omega\hat{c}_\varepsilon(x,\lambda)A^\varepsilon(x)\nabla u(x)\overline{\nabla v}(x)\,dx \tag{5.9}$$

is continuous coercive on $H_0^1(\Omega;\mathbb{C})$ under the hypothesis

$$\operatorname{Re}\hat{c}_\varepsilon(x,\lambda)\geqq\rho_0(\lambda)>0, \tag{5.10}$$

which we will check later.

Indeed, c_ε is bounded on Ω for each λ such that $\operatorname{Re}\lambda>0$ so a_ε is continuous and for such λ's,

$$\operatorname{Re}a_\varepsilon(\lambda;u,u)=\int_\Omega\operatorname{Re}\hat{c}_\varepsilon(x,\lambda)A^\varepsilon(x)|\nabla u(x)|^2\,dx$$

$$\geqq\alpha\int_\Omega\operatorname{Re}\hat{c}_\varepsilon(x,\lambda)|\nabla u|^2\,dx$$

since A^ε is symmetric real coercive. Incidentally, another proof of existence for the solution u^ε is thus obtained by applying Lax–Milgram's theorem for $\hat{u}^\varepsilon$.

One can now apply a compactness result for complex homogenization (see for example Sanchez-Palencia [1], Murat [1] or Bensoussan–Lions–Papanicolaou [1]). For each λ with $\mathrm{Re}\lambda > \lambda_0$, there is a matrix-valued function $D(x,\lambda)$ (independent of $\hat{\mathscr{F}}$ and $\hat{\mathscr{W}}$) such that $(5.8)_\varepsilon$ implies at the limit $\varepsilon \to 0$:

$$\begin{aligned} &-\operatorname{div}\hat{\mathscr{W}} = \hat{\mathscr{F}}, \\ &\hat{\mathscr{W}}(x,\lambda) = D(x,\lambda)\nabla\hat{u}(x,\lambda). \end{aligned} \tag{5.11}$$

In order to apply the inverse Laplace transform to (5.11), all that is needed is that $D(x,\lambda)$ be analytic in λ with at most polynomial growth at $|\lambda| \to \infty$, in which case it is the Laplace transform of a distribution of finite order in t, denoted $E(x,t)$. That D is analytic is a mere consequence of the fact that it is a limit of a sequence of analytic functions of λ, the limit being locally uniform. From the uniform boundedness of c^ε and A^ε, one can conclude that $D(x,\lambda)$ is bounded by a multiple of $(\mathrm{Re}\lambda)^{-1}$. Consequently, $E(x,t)$ is a bounded distribution of order not more than 2, on $[0, +\infty[$, with values in the cone of bounded measurable symmetric square matices on Ω.

We now check that (5.10) holds: integration by parts in

$$\mathrm{Re}\,\hat{c}_\varepsilon(x,\lambda) = \int_0^\infty e^{-(\mathrm{Re}\lambda)t}\cos(\mathrm{Im}\lambda)\,c_\varepsilon(t)\,dt \tag{5.12}$$

gives

$$\mathrm{Re}\,\hat{c}_\varepsilon(x,\lambda) = \int_0^\infty (1-\cos(t\,\mathrm{Im}\lambda))\,e^{-t\,\mathrm{Re}\lambda}\left(c_\varepsilon'' - 2\,\mathrm{Re}\lambda\,c_\varepsilon' + (\mathrm{Re}\lambda)^2 c_\varepsilon\right)dt.$$

Since $(1-\cos(t\,\mathrm{Im}\lambda))$ and $-c_\varepsilon'$ are nonnegative functions and c_ε'' is a nonnegative measure,

$$\mathrm{Re}\,\hat{c}_\varepsilon(x,\lambda) \geqq \int_0^{+\infty} (1-\cos(t\,\mathrm{Im}\lambda))\,e^{-t\,\mathrm{Re}\lambda}c_\varepsilon(t)\,dt; \tag{5.13}$$

combining (5.12) and (5.13) one gets

$$\mathrm{Re}\,\hat{c}_\varepsilon(x,\lambda) \geqq \frac{(\mathrm{Re}\lambda)^2}{1+(\mathrm{Re}\lambda)^2}\int_0^\infty e^{-t\,\mathrm{Re}\lambda}c_\varepsilon(t)\,dt. \tag{5.14}$$

But $c_\varepsilon(t) \geqq \alpha$ implies with (5.14) that $\mathrm{Re}\,\hat{c}_\varepsilon(x,\lambda) \geqq \alpha\mathrm{Re}\lambda/(1+(\mathrm{Re}\lambda)^2)$ which implies (5.10). Finally, we have proved the following theorem:

THEOREM 5.15. *Let b_0^ε, c^ε, β^ε and $\sigma^\varepsilon = A^\varepsilon$ (linear case) satisfy hypotheses* (4.4) *with α and k independent of ε. There exists a sequence ε_n converging to zero, functions $b_0(x)$, $\beta(x,t)$ and a distribution $E(x,t)$ such that the solution u^{ε_n} of the corresponding problem* (4.1), (4.2) *converges in $C([0,T];L^2(\Omega))$ to the solution u of*

$$b_0 u' - \operatorname{div}\left((E * \nabla u)'\right) = G(u). \tag{5.16}$$

b_0 and β are the weak limits of $b_0^{\varepsilon_n}$ and β^{ε_n}, and E is obtained via its Laplace transform $D(x,\lambda)$, which is the complex elliptic homogenization of $\hat{c}_{\varepsilon_n}(x,\lambda)A^{\varepsilon_n}(x)$.

Remark. Even in the case of periodic problems, where there are explicit formulas for D it is not known whether $t \mapsto E(t,x)$ is in some appropriate sense, a convex decreasing function of t, not even whether it is a function of t, as one would suspect. This is one of the problems left open in the theory, the other one being the homogenization of the nonlinear case (where the Laplace-transform cannot be used).

Acknowledgment. The authors are very grateful to Professor John Nohel who not only introduced them to nonlinear Volterra equations in general and this problem in particular, but without whose help this paper never would have been written.

REFERENCES

H. ATTOUCH [1], *Sur la Γ-convergence. Nonlinear partial differential equations and their applications*, Collège de France Seminar, Vol. I, H. Brezis and J. L. Lions, eds., Pitman, London, 1980, pp. 7–41.

_____ [2], *Variational convergences for functions and operators*, in Applicable Mathematics, Pitman, London, 1984.

H. ATTOUCH AND A. DAMLAMIAN [1], *Application des méthodes de convexité et monotonie à l'étude de certaines équations quasi-linéaires*, Proc. Royal Soc. Edinburgh, 79A (1977), pp. 107–129.

J. P. AUBIN [1], *Un théorème de compacité*, CRAS Paris, 256A (1963), pp. 5042–5044.

A. BENSOUSSAN, J. L. LIONS AND G. PAPANICOLAOU [1], *Asymptotic analysis for periodic structures*, in Studies in Mathematics and Its Applications, North-Holland, Amsterdam, 1978.

H. BREZIS [1], *Perturbations non linéaires d'opérateurs maximaux monotones*, CRAS Paris, 269A (1969), pp. 566–569.

M. CRANDALL AND J. NOHEL [1], *An abstract functional differential equation and a related nonlinear Volterra equation*, Israel J. Math., 29 (1978), pp. 313–328.

U. MOSCO [1], *Convergence of convex set and of solutions of variational inequalities*, Adv. Math., 3 (1969), pp. 510–585.

F. MURAT [1], *H-convergence. Séminaire d'analyse functionnelle et numérique*, Université d'Alger, 1977–78.

J. NOHEL [1], *Nonlinear Volterra equations for heat flow in materials with memory*, Technical Summary Report 208, Mathematics Research Center, University of Wisconsin-Madison, Madison, WI, 1980.

M. L. RAYNAL [1], CRAS Paris, 292.1 (1981), pp. 421–424.

E. SANCHEZ–PALENCIA [1], *Nonhomogeneous media and vibration theory*, Lecture Notes in Physics 127, Springer, Berlin, 1980.

SIAM J. MATH. ANAL.
Vol. 17, No. 6, November 1986

011

A NONLINEAR SINGULAR INTEGRO-DIFFERENTIAL EQUATION ARISING IN SURFACE CHEMISTRY*

JEAN DUCHON† AND RAOUL ROBERT†

Abstract. We prove an existence-uniqueness result for the solution of a nonlinear singular integro-differential equation. This equation is an approximate model for the development, by the mechanism of volume diffusion, of a grain boundary groove on an interface separating a solid phase and a saturated fluid phase.

Key words. nonlinear integro-differential equation, special equations and problems

AMS(MOS) subject classification. Primary 35Q

Introduction. When a polycrystalline solid is heated in thermodynamical equilibrium with a fluid phase, grooves will form at the intersections of the grain boundaries and the interphase interface. These grooves have a constant dihedral angle and their dimensions roughly vary as $(\text{time})^{\alpha}$, the exponent α being determined by the dominant transport process involved.

We study here a model for the development by the mechanism of volume diffusion of such a grain boundary groove.

Referring to Mullins [6] for more details, we give a brief summary of the most important physical assumptions upon which the model is based:

1) the transport process involved in the growth of the groove is volume diffusion in the fluid phase,
2) isotropy of the interfacial free energy,
3) applicability of the Gibbs–Thompson formula relating curvature and chemical potential,
4) quasi steady-state volume diffusion,
5) negligible convection in the fluid phase,
6) the interface is taken to be initially flat.

Writing up the equations governing the groove development, we show that the groove profile has a fixed shape and linear dimensions that are proportional to $(\text{time})^{1/3}$.

This led us to study the stationary equations giving the shape of the groove. Then, we make some sort of nonlinear small slope approximation. That is, letting the curvature term be unchanged, for purposes of the diffusion problem we represent the interface as a plane.

The study of this approximate problem is a first step towards the solution of the complete problem (which is object of actual research), it gives some insight on the way the operators involved behave. On the same way the numerical resolution of the approximate problem (see [7]) enlightens our understanding of the computational processes involved in the numerical resolution of the complete problem.

From a physical point of view, the practical value of such an approximation is not obvious. Roughly speaking, the correction term created by our approximation depends linearly on the slope of the groove root (this is discussed on a physical level in [6]) whereas the correction term given by the linearization of the curvature depends

*Received by the editors December 6, 1983, and in revised form January 11, 1985.
†Laboratoire I.M.A.G., B. P. 68, 38402 Saint-Martin-d'Hères Cedex, France.

quadratically on the slope. So, for small slopes, we cannot expect a better suited quantitative model than the completely linear small slope approximation given in [6].

1. The mathematical problem and its approximation. We consider a plane section normal to the groove line and define cartesian coordinates in the plane so that the x axis coincides with the trace of the initially flat interface and the negative y axis coincides with the grain boundary. Letting $y=w(t,x)$ be the profile of the groove, we note that

$$\Omega_t=\{(x,y)\,|\,y>w(t,x)\},\qquad \partial\Omega_t=\{(x,y)\,|\,y=w(t,w)\},$$

$n(x,w(t,x))$ the unit vector normal to $\partial\Omega_t$ at $(x,w(t,x))$ and pointing out of Ω_t, $C(t,x,y)$ the concentration of solid atoms in the solvent.

Then, following Mullins [6], we write the set of equations governing the growth of the groove:

(i) $\Delta C=0$ in Ω_t (quasi steady-state diffusion),

(ii) $C=\mathrm{Co}(1-L(D^2w/(1+Dw^2)^{3/2}))$ on $\partial\Omega_t$ for $x\neq 0$ (Gibbs–Thompson formula),

(iii) $Dw(t,0\pm)=\pm m$ for all $t>0$ (constant dihedral angle),

(iv) $\partial w/\partial t=-M(1+Dw^2)^{1/2}\partial C/\partial n$ (the normal flux determines the rate of movement of the interface),

(v) $w(0,x)=0$ (initially flat interface),

where m is the slope at the groove root, D indicates differentiation with respect to x and Co, L, M are physical constants.

We seek a solution of constant shape and try the function change:

$$w(t,x)=t^{1/3}u(t^{-1/3}x),$$
$$C(t,x,y)=\mathrm{Co}\big(1-t^{-1/3}c\big(t^{-1/3}x,t^{-1/3}y\big)\big).$$

One readily sees by a straightforward computation that equations (i) $\cdots$ (v) change into the following system of stationary equations.

$$\text{(S)}\qquad \begin{aligned} &\Delta c=0 \quad \text{in } \Omega,\\ &c=L\frac{D^2u}{(1+Du^2)^{3/2}} \quad \text{on } \partial\Omega \text{ for } x\neq 0,\\ &Du(0\pm)=\pm m,\\ &u-xDu=3\,\mathrm{Co}M(1+Du^2)^{1/2}\frac{\partial c}{\partial n}, \end{aligned}$$

where $\Omega=\{(x,y)\,|\,y>u(x)\}$.

Let us identify c on $\partial\Omega$ with a function of x, i.e. $c(x)=c(x,u(x))$. We show in [4], by a study of the single layer potential, the splitting:

$$(1+Du^2)^{1/2}\frac{\partial c}{\partial n}=2\pi\Lambda^1 c+P(u)\cdot Dc,$$

where the operator Λ^s is defined by Fourier transform

$$\widehat{\Lambda^s f}(\xi)=|\xi|^s\hat{f}(\xi).$$

When u is a smooth function the operator $P(u)$ (taken on the space $L^2(\mathbb{R})$) is a compact and smoothing one whose norm, denoted $\|P(u)\|_{op}$, goes to 0 with u (in some convenient norm). When u is only Lipschitz $P(u)$ is defined by means of a singular

kernel and we can see, using deep results of Coifman, McIntosh and Meyer [1] that $\|P(u)\|_{op}$ goes to 0 as the Lipschitz constant of u goes to 0.

For a detailed study of operator $P(u)$, which falls out of the scope of this work, we refer to [4] and [1].

In the sequel we shall approxime the operator $(1+Du^2)^{1/2}\partial/\partial n$ by its "dominant part" $2\pi\Lambda^1$.

Then (S) becomes

$$\text{(E)} \qquad u - xDu - \gamma\Lambda^1 \frac{[D^2u]}{(1+Du^2)^{3/2}} = 0, \qquad Du(0\pm) = \pm m,$$

and we note $[D^2u] = D^2u - 2m\delta$, where δ is the Dirac mass in 0 and derivatives are taken in the distribution sense.

If we approximate the curvature of the interface by D^2u, (E) becomes the linear small slope approximation studied in [6].

By a change of scale in the space, we can take the constant γ to be $1/4\pi^2$ which will be a convenient choice for the sequel.

We prove here an existence and uniqueness result for the solution of equation (E). Let us briefly indicate the plan of the proof.

First step (paragraph 2). For $\varepsilon > 0$, we introduce an elliptic regularization (E_ε) of equation (E). Using a suitable linearization of (E_ε) and two estimates (A1, A2), we apply Schauder's fixed point theorem to prove that (E_ε) has a solution (Theorem 2.1).

Second step (paragraph 3). We show that equation (E) has a unique solution (Theorem 3.1) by a limit process, letting ε go to zero in (E_ε). Here the crucial point is to obtain the key estimate $|Du_\varepsilon|_\infty$ bounded. This is a consequence of some technical lemmas and three basic estimates on u_ε solution of (E_ε), (B1, B2, B3). Uniqueness is a straightforward consequence of the monotony of the nonlinear operator $-D^2u/(1+Du^2)^{3/2}$.

2. The regularized equation (E_ε). Throughout this paper we note $\operatorname{sgn}\xi$ the sign of the real number ξ and $\operatorname{Re} z$ the real part of the complex number z. Functions under consideration are real-valued. Notice that for f real valued, $\Lambda^s f$ is real valued. For s real, we note $H^s = H^s(\mathbb{R})$ the usual Sobolev's space and $\langle \cdot, \cdot \rangle$ the pairing between H^s and H^{-s}. Also we note $|f|$ the L^2-norm and $|f|_\infty$ the L^∞-norm of the function f.

We find it convenient to use the solution u_0 of the linear problem:

$$u_0 - xDu_0 - \frac{1}{4\pi^2}\Lambda^1[D^2u_0] = 0,$$
$$Du_0(0\pm) = \pm m$$

or equivalently:

$$u_0 - xDu_0 - \frac{1}{4\pi^2}\Lambda^1 D^2 u_0 = -\frac{m}{2\pi^2}\Lambda^1\delta.$$

Fourier transformation leads to the linear differential equation

$$\hat{u}_0 + D(\xi\hat{u}_0) + |\xi|^3\hat{u}_0 = -6m|\xi|,$$

and standard calculation gives the unique tempered solution:

$$\hat{u}_0(\xi) = -\frac{m}{2\pi^2\xi^2}\left(1 - e^{-|\xi|^3/3}\right).$$

We notice that u_0 is in H^1 and

$$D^2u_0=[D^2u_0]+2m\delta$$

with $[D^2u_0]$ square integrable.

We then introduce, for $\varepsilon>0$, the regularized equation

$$(E_\varepsilon)\qquad \begin{aligned} &u-xDu-\frac{1}{4\pi^2}\Lambda^1\frac{[D^2u]}{(1+Du^2)^{3/2}}-\frac{\varepsilon}{4\pi^2}\Lambda^1(D^2u-D^2u_0)=0,\\ &Du(0\pm)=\pm m. \end{aligned}$$

In this section we prove the following:

THEOREM 2.1. *Equation* (E_ε) *possesses a unique solution* u_ε *in the set* u_0+H^2. *This function is even.*

Proof. Let $u=u_0+v$, $v\in H^2$, and define

$$T(g)=\left(1-\left(1+(Du_0+g)^2\right)^{-3/2}\right)(Dg+[D^2u_0]);$$

equation (E_ε) now is:

$$\begin{aligned} &v-xDv+(1+\varepsilon)\Lambda^3v=-\frac{1}{4\pi^2}\Lambda^1T(Dv),\\ &Dv(0)=0. \end{aligned}$$

For $f\in L^2(\mathbb{R})$, we define $S(f)$ to be the unique solution $v\in H^2$ of

$$(LE_\varepsilon)\qquad v-xDv+(1+\varepsilon)\Lambda^3v=-\frac{1}{4\pi^2}\Lambda^1f.$$

Equation (E_ε) then can be written

$$\begin{aligned} &v=S(T(Dv))=\Phi(v),\\ &Dv(0)=0. \end{aligned}$$

We show now that the equation $v=\Phi(v)$ possesses an even solution in H^2. For $r\geqq 0$, let C_r be the weakly compact convex set in H^2 of even functions v such that: $|v|$, $|Dv|$, $|D^2v|\leqq r$. for r large enough, Φ maps C_r into itself as is shown by the following lemma:

LEMMA 2.2. *Let* $f\in L^2(\mathbb{R})$ *and* $v=S(f)$. *Then we have*:

(i) $|v|\leqq c|f|$,
(ii) $|Dv|\leqq c|f|$,
(iii) $|Dv|_\infty\leqq c|f|$,
(iv) $|D^2v|\leqq(1+\varepsilon)^{-1/2}|f|$,

where c *is a constant* <1.

Proof. By Fourier transformation equation (LE_ε) becomes

$$\hat v+D(\xi\hat v)+(1+\varepsilon)|\xi|^3\hat v=-\frac{1}{4\pi^2}|\xi|\hat f$$

whose unique tempered solution is given by:

$$(0)\qquad \hat v(\xi)=\frac{-1}{4\pi^2\xi^2}e^{-(1+\varepsilon)|\xi|^3/3}\int_0^\xi e^{(1+\varepsilon)|s|^3/3}s|s|\hat f(s)\,ds.$$

Schwarz's inequality shows this function is locally square integrable, since

$$\left|\int_0^\xi e^{(1+\varepsilon)|s|^3/3}s|s|\hat{f}(s)\,ds\right| \leqq \left|\int_0^\xi |\hat{f}(s)|^2\,ds\right|^{1/2}\cdot\left|\int_0^\xi s^4 e^{2(1+\varepsilon)|s|^3/3}\,ds\right|^{1/2}$$

but

$$\left|\int_0^\xi s^4 e^{2(1+\varepsilon)|s|^3/3}\,ds\right| \leqq |\xi|^5 e^{2(1+\varepsilon)|\xi|^3/3}$$

so that $|\hat{v}(\xi)| \leqq (1/4\pi^2)|\xi|^{1/2}|\hat{f}|$.

But it is not obvious from the expression (0) of $\hat{v}$ above that $v \in H^2$. Therefore we shall show inequalities (i),$\cdots$,(iv) for $\hat{f}$ C^∞ with compact support, a density argument then completes the proof of the lemma.

Estimate (A1). Multiplying equation (LE_ε) by v and integrating over $\mathbb{R}$, we get

$$|v|^2 - \int x\,Dv\,v\,dx + (1+\varepsilon)|\Lambda^{3/2}v|^2 \leqq \frac{1}{4\pi^2}|f\,||\Lambda^1 v|.$$

One easily verifies using formula (0) that this makes sense. Integration by parts gives

$$-\int x\,Dv\,v\,dx = \int D(xv)\,v\,dx = \int v^2\,dx + \int xv\,Dv\,dx,$$

whence $-\int x\,Dv\,v\,dx = \frac{1}{2}|v|^2$, and we get

$$\text{(A1)} \qquad \frac{3}{2}|v|^2 + (1+\varepsilon)|\Lambda^{3/2}v|^2 \leqq \frac{1}{4\pi^2}|f\,||\Lambda^1 v|.$$

Estimate (A2). We multiply equation (LE_ε) by $\Lambda^1 v$ and integrate over $\mathbb{R}$; it becomes

$$|\Lambda^{1/2}v|^2 - \int x\,Dv\Lambda^1 v\,dx + (1+\varepsilon)|\Lambda^2 v|^2 \leqq \frac{1}{4\pi^2}|f\,||\Lambda^2 v|.$$

Notice that $-\int x\,Dv\Lambda^1 v\,dx = \int D(\xi\hat{v})|\xi|\bar{\hat{v}}\,d\xi = 0$; then we get

$$\text{(A2)} \qquad |\Lambda^{1/2}v|^2 + \frac{1+\varepsilon}{2}|\Lambda^2 v|^2 \leqq \frac{1}{2(4\pi^2)^2}|f|^2$$

which proves (iv):

$$(1+\varepsilon)^{1/2}|D^2 v| \leqq |f|.$$

We now use estimates (A1) and (A2) to prove (i), (ii), (iii). An easy computation gives

$$|Dv|^2 \leqq \frac{8\pi^2}{3}\left(|\Lambda^{1/2}v|^2 + \frac{1}{2}|\Lambda^2 v|^2\right),$$

which together with (A2) gives

$$|Dv|^2 \leqq \frac{1}{12\pi^2}|f|^2,$$

i.e., (ii).

Using (ii), (iv) and $|Dv|_\infty \leqq 1/\sqrt{2}\,(|Dv|^2+|D^2v|^2)^{1/2}$ one gets

$$|Dv|_\infty \leqq \left(\frac{1}{2}+\frac{1}{24\pi^2}\right)^{1/2}|f|,$$

i.e., (iii).

Finally (i) is deduced from (A1) and (ii). □

One easily sees now that Φ maps c_r into itself for large r. First $\Phi(v)$ is even if v is even, and Lemma 2.2 says:

$$|\Phi(v)|, |D\Phi(v)|, |D^2\Phi(v)| \leqq c_\varepsilon |T(Dv)|$$

with $c_\varepsilon = \max(c, (1+\varepsilon)^{-1/2}) < 1$.

But $|T(Dv)| \leqq |D^2v| + |[D^2u_0]|$, so that r just has to be chosen large enough to ensure:

$$c_\varepsilon\left(r + |[D^2u_0]|\right) \leqq r.$$

Let us now prove some convenient continuity property for the operator Φ.

LEMMA 2.3. *The operator T is continuous from bounded subsets of H^1 into L^2 for the weak topologies.*

Proof. Recall $T(f) = (1-(1+(Du_0+f)^2)^{-3/2})(Df+[D^2u_0])$. Let f_n converge weakly to f in H^1. Then $f_n \to f$ uniformly on every compact and $Df_n \to Df$ weakly in L^2. An easy extra lemma says if u_n is bounded in L^∞ and $u_n \to u$ uniformly on every compact and $v_n \to v$ weakly in L^2 then $u_n v_n \to uv$ weakly in L^2. This, applied to $(1+(Du_0+f_n)^2)^{-3/2}$ and Df_n completes the proof. □

The linear operator S is continuous from L^2 into H^2, then also weakly continuous, and from Lemma 2.3, one concludes Φ is continuous on C_r for the weak topology of H^2. By Schauder's theorem, Φ has a fixed point $v_\varepsilon \in C_r$. Letting $u_\varepsilon = u_0 + v_\varepsilon$, one gets an even solution of equation (E_ε). This solution is unique (the proof of uniqueness is as in the following section in the case of equation (E)). □

3. An existence-uniqueness result for equation (E). This section is devoted to proving the following:

THEOREM 3.1. *Equation* (E) *has a unique solution u in the set $u_0 + H^2$. This solution is even.*

Proof. We let ε go to zero in equation (E_ε). Before deriving the necessary a priori estimates, we give a technical lemma.

LEMMA 3.2. *Let $h \in L^2$. There is a unique tempered function u solution of*

$$u - x\,Du = \Lambda^1 h \tag{1}$$

satisfying the estimate

$$|\Lambda^{-1}u|^2 + 3\pi^2|xu|^2 \leqq |h|^2. \tag{2}$$

If moreover $u \in L^2$, then

$$\Lambda^{-1/2}u \in L^2 \quad \text{and} \quad \langle h, u\rangle = 2|\Lambda^{-1/2}u|^2. \tag{3}$$

Proof. Applying Fourier transforms to (1) gives

$$2\hat{u} + \xi D\hat{u} = |\xi|\hat{h}; \tag{$\hat{1}$}$$

this differential equation has for particular solution

$$\hat{u}(\xi)=\xi^{-2}\int_0^{\xi} s|s|\hat{h}(s)\,ds.$$

Applying Schwarz's inequality, one sees that $|\hat{u}(\xi)|\leqq C|\xi|^{1/2}$, so $\hat{u}$ is a tempered function.

Now we show this particular solution satisfies estimate (2): first considering the case $\hat{h}$ C^∞ with compact support, the result will then follow for $h\in L^2$ by a density argument.

For such a function h, $\xi^{-1}\hat{u}$ and $D\hat{u}$ both are in L^2, and so that we can multiply equation $(\hat{1})$ by $\xi^{-1}D\bar{\hat{u}}$ and integrate over $\mathbb{R}$, we get:

$$2\int\hat{u}\,\frac{D\bar{\hat{u}}}{\xi}\,d\xi+\int|D\hat{u}|^2\,d\xi=\int\operatorname{sgn}(\xi)\hat{h}\,D\bar{\hat{u}}\,d\xi;$$

u being real, the first integral is real and we have:

$$\int\hat{u}\,D\bar{\hat{u}}\xi^{-1}\,d\xi=\int\xi^{-1}\operatorname{Re}(\hat{u}\,D\bar{\hat{u}})\,d\xi=\frac{1}{2}\int\xi^{-1}D|\hat{u}|^2\,d\xi;$$

an integration by parts performed on the last integral then gives

$$\int\hat{u}\,D\bar{\hat{u}}\xi^{-1}\,d\xi=\frac{1}{2}\int\xi^{-2}|\hat{u}|^2\,d\xi.$$

Then we have

$$|\Lambda^{-1}u|^2+4\pi^2|xu|^2\leqq 2\pi|h||xu|\leqq|h|^2+\pi^2|xu|^2,$$

which proves estimate (2).

Uniqueness is obvious here since the general solution of (1) is obtained by adding λx to the particular solution.

If u is square integrable, multiplying equation $(\hat{1})$ by $|\xi|^{-1}\bar{\hat{u}}$ gives

$$2\int\hat{u}\bar{\hat{u}}|\xi|^{-1}\,d\xi+\int\operatorname{sgn}(\xi)\,D\hat{u}\,\bar{\hat{u}}\,d\xi=\int\hat{h}\bar{\hat{u}}\,d\xi.$$

But

$$\int\operatorname{sgn}(\xi)\,D\hat{u}\,\bar{\hat{u}}\,d\xi=\int\operatorname{sgn}(\xi)\,\operatorname{Re}(D\hat{u}\bar{\hat{u}})\,d\xi=\frac{1}{2}\int\operatorname{sgn}(\xi)\,D|\hat{u}|^2\,d\xi,$$

and this last integral vanishes since $\hat{u}$ is zero at 0 and ∞ (notice $\hat{u}\in H^1$). Then we have $\langle h,u\rangle=2\int|\hat{u}|^2|\xi|^{-1}\,d\xi$, i.e. (3). □

We use Lemma 3.2 to prove the following estimates.

Estimate (B1). Equation (E_ε) writes $u-xDu=\Lambda^1 h$,[1] with

$$h=\frac{1}{4\pi^2}\left(\frac{[D^2u]}{(1+Du^2)^{3/2}}+\varepsilon D^2u-\varepsilon D^2u_0\right).$$

Lemma 3.2 then gives

$$\int\frac{[D^2u]u}{(1+Du^2)^{3/2}}\,dx+\varepsilon\langle D^2u,u\rangle-\varepsilon\langle D^2u_0,u\rangle=8\pi^2|\Lambda^{-1/2}u|^2.$$

[1] For simplicity we denote by u instead of u_ε the solution of equation (E_ε).

Integration by parts gives

$$\int \frac{[D^2u]u}{(1+Du^2)^{3/2}}\,dx = -\int \frac{Du^2}{(1+Du^2)^{1/2}}\,dx - \frac{2m}{(1+m^2)^{1/2}}u(0);$$

we have also

$$\langle D^2u_0, u\rangle = \langle [D^2u_0],\ u\rangle + 2mu(0),$$

and then

$$\text{(B1)}\quad 8\pi^2|\Lambda^{-1/2}u|^2 + \int \frac{Du^2}{(1+Du^2)^{1/2}}\,dx + \varepsilon|Du|^2 + 2m\big((1+m^2)^{-1/2}+\varepsilon\big)u(0) = -\varepsilon\langle [D^2u_0],\ u\rangle.$$

Estimate (B2). We may write equation (E_ε) as follows:

$$\Lambda^{-1}(u-xDu) - \frac{1}{4\pi^2}\frac{[D^2u]}{(1+Du^2)^{3/2}} - \frac{\varepsilon}{4\pi^2}(D^2u - D^2u_0) = 0.$$

We multiply by $-[D^2u]$ and integrate over $\mathbb{R}$. Then we need the following lemma.

LEMMA 3.3. *The solution u of* (E_ε) *satisfies*:

$$\langle \Lambda^{-1}(u-xDu),\ -[D^2u]\rangle = 4\pi^2|\Lambda^{1/2}u|^2 + 4m\Lambda^{-1}u(0).$$

Proof. We first notice that $|\xi|^{-1}\hat{u}$ is integrable, so that $\Lambda^{-1}u$ is a continuous function (vanishing to infinity) and $\Lambda^{-1}u(0)$ makes sense. Let us begin the proof with a formal computation where some of the scalar products may be undefined. One has

$$D^2u = [D^2u] + 2m\delta,$$

and then

$$\langle \Lambda^{-1}(u-xDu),\ -[D^2u]\rangle = \langle \Lambda^{-1}(u-xDu),\ -D^2u\rangle + 2m\langle \Lambda^{-1}(u-xDu),\ \delta\rangle,$$

but

$$\langle \Lambda^{-1}(u-xDu),\ -D^2u\rangle = 4\pi^2\langle u-xDu,\ \Lambda^1u\rangle = 4\pi^2|\Lambda^{1/2}u|^2 - 4\pi^2\langle xDu,\ \Lambda^1u\rangle$$

and

$$\langle \Lambda^{-1}(u-xDu),\ \delta\rangle = \langle \Lambda^{-1}(2u-D(xu)),\ \delta\rangle = 2\Lambda^{-1}u(0) - \langle \Lambda^{-1}D(xu),\ \delta\rangle.$$

Now we examine the term

$$\langle \Lambda^{-1}D(xu),\ \delta\rangle = -\int \operatorname{sgn}(\xi)\,D\hat{u}\,d\xi.$$

This last integral might be undefined since $D\hat{u}$ is only square integrable; however $\lim_{A\to\infty}\int_{-A}^{+A}\operatorname{sgn}(\xi)\,D\hat{u}d\xi = 0$ for $\hat{u}$ is in H^1 and $\hat{u}(0)=0$. Similarly $\langle xDu,\ \Lambda^1u\rangle = -\int D(\xi\hat{u})|\xi|\bar{\hat{u}}\,d\xi$.

Let us prove that

$$\lim_{A\to+\infty} \operatorname{Re}\int_{-A}^{+A} \operatorname{sgn}(\xi) D(\xi\hat{u})\xi\bar{\hat{u}}\,d\xi = 0,$$

$$\operatorname{Re}\int_{-A}^{+A} \operatorname{sgn}(\xi) D(\xi\hat{u})\xi\bar{\hat{u}}\,d\xi = \frac{1}{2}\int_{-A}^{+A} \operatorname{sgn}(\xi) D\left(\xi^2|\hat{u}|^2\right) d\xi.$$

Thus it is enough to prove that $\xi\hat{u}$ vanishes to infinity, which is a consequence of the following lemma.

LEMMA 3.4. *Let the function v satisfy a Hölder condition of order $\alpha/2$, $0<\alpha\leqq 2$. If $\xi^{1+\alpha}v$ is square integrable, then $\xi^{\alpha}v\to 0$ as $|\xi|\to\infty$.*

Proof. By contradiction. Thus assume the existence of a sequence ξ_n, $|\xi_n|\to\infty$, and $\varepsilon>0$ such that $|v(\xi_n)||\xi_n|^{\alpha}\geqq\varepsilon$. Since v satisfies a Hölder condition of order $\alpha/2$,

$$|v(\xi)-v(\xi_n)|\leqq c|\xi-\xi_n|^{\alpha/2};$$

therefore

$$|v(\xi)|\geqq|v(\xi_n)|-c|\xi-\xi_n|^{\alpha/2}.$$

Call In the interval $[\xi_n-\lambda\xi_n^{-2},\ \xi_n+\lambda\xi_n^{-2}]$, $\lambda>0$. For $\xi\in$ In, we have $|\xi-\xi_n|^{\alpha/2}\leqq\lambda^{\alpha/2}|\xi_n|^{-\alpha}$, so that a suitable choice of λ yields $|v(\xi)|\geqq\varepsilon/2|\xi_n|^{-\alpha}$.

If the In's are disjoint, which we may assume, one has

$$\int|\xi|^{2\alpha+2}|v|^2\,d\xi\geqq\sum_{n=0}^{\infty}\int_{\mathrm{In}}|\xi|^{2\alpha+2}|v|^2\,d\xi,$$

but

$$\int_{\mathrm{In}}|\xi|^{2\alpha+2}|v|^2\,d\xi\geqq\lambda\varepsilon^2/4,$$

so that the integral is divergent, which proves Lemma 3.4. □

Let us go back to the proof of Lemma 3.3.

$$\hat{u}=\hat{u}_0+\hat{v}\quad\text{and}\quad\xi\hat{u}_0=\frac{m}{2\pi^2\xi}\left(e^{-|\xi|^3/3}-1\right).$$

Thus it is enough to prove that $\xi\hat{v}$ vanishes to infinity. $\hat{u}$ and $\hat{u}_0$ belong to the space H^1; so do $\hat{v}$, which then satisfies a Hölder condition of order $\frac{1}{2}$, and applying Lemma 3.4 with $\alpha=1$ yields the desired result.

Finally, to justify the formal computation in Lemma 3.3, we compute $\left\langle\Lambda^{-1}(u-xDu),\ K_A*[D^2u]\right\rangle$, where $\hat{K}_A$ is the characteristic function of $[-A,+A]$, and take the limit $A\to\infty$. This completes the proof of Lemma 3.3. □

From Lemma 3.3 one gets the estimate

$$\text{(B2)}\qquad 4\pi^2|\Lambda^{1/2}u|^2+\frac{1}{4\pi^2}\int\frac{[D^2u]^2}{(1+Du^2)^{3/2}}\,dx+\frac{\varepsilon}{8\pi^2}\left|[D^2u]\right|^2$$

$$\leqq\frac{\varepsilon}{8\pi^2}\left|[D^2u_0]\right|^2-4m\Lambda^{-1}u(0).$$

Estimate (B3). We proceed as above but multiplying by $\Lambda^{-1}u$. Let us compute the term:

$$\langle \Lambda^{-1}(u-xDu),\ \Lambda^{-1}u\rangle = 2|\Lambda^{-1}u|^2 - \langle \Lambda^{-1}D(xu),\ \Lambda^{-1}u\rangle$$

but

$$-\langle \Lambda^{-1}D(xu),\ \Lambda^{-1}u\rangle = \int D\hat{u}\bar{\hat{u}}\xi^{-1}\,d\xi,$$

and, as in the proof of Lemma 3.2, we see that

$$\int D\hat{u}\bar{\hat{u}}\xi^{-1}\,d\xi = \frac{1}{2}|\Lambda^{-1}u|^2,$$

and therefore

$$\langle \Lambda^{-1}(u-xDu),\ \Lambda^{-1}u\rangle = \frac{5}{2}|\Lambda^{-1}u|^2.$$

Denoting

$$Bu = -\frac{1}{4\pi^2}\frac{[D^2u]}{(1+Du^2)^{3/2}},$$

one obtains

$$\frac{5}{2}|\Lambda^{-1}u|^2 + \langle Bu,\ \Lambda^{-1}u\rangle - \frac{\varepsilon}{4\pi^2}\langle D^2u,\ \Lambda^{-1}u\rangle + \frac{\varepsilon}{4\pi^2}\langle D^2u_0,\ \Lambda^{-1}u\rangle = 0.$$

Then

$$\frac{5}{2}|\Lambda^{-1}u|^2 + \varepsilon|\Lambda^{1/2}u|^2 \leqq |Bu||\Lambda^{-1}u| + \varepsilon|\Lambda^{1/2}u_0||\Lambda^{1/2}u|,$$

and finally

$$2|\Lambda^{-1}u|^2 + \frac{\varepsilon}{2}|\Lambda^{1/2}u|^2 \leqq \frac{\varepsilon}{2}|\Lambda^{1/2}u_0|^2 + \frac{1}{2}|Bu|^2. \tag{B3}$$

We now use estimates (B1), (B2), (B3) to prove Theorem 3.1. First, for every $\alpha > 0$ there is a $C\alpha \geqq 0$ such that

$$|\Lambda^{-1}u|_\infty \leqq \alpha|\Lambda^{1/2}u| + C\alpha|\Lambda^{-1}u|.$$

Taking this inequality in (B2) with $\alpha = \pi^2/m$ yields

$$2\pi^2|\Lambda^{1/2}u|^2 + \frac{1}{4\pi^2}\int \frac{[D^2u]^2}{(1+Du^2)^{3/2}}\,dx + \frac{\varepsilon}{8\pi^2}|[D^2u]|^2 \leqq A_1 + A_2|\Lambda^{-1}u|$$

for some constants A_1 and A_2.

On the other hand, we derive from (B3)

$$\sqrt{2}\,|\Lambda^{-1}u| \leqq \sqrt{\frac{\varepsilon}{2}}\,|\Lambda^{1/2}u_0| + \frac{1}{\sqrt{2}}|Bu|,$$

which together with the previous inequality and

$$\int \frac{[D^2u]^2}{(1+Du^2)^{3/2}}\,dx \geqq 16\pi^4|Bu|^2$$

gives $4\pi^2|Bu|^2 \leqq A_3 + A_4|Bu|$.

From this we deduce that $|Bu_\varepsilon|$ remains bounded and, going back to (B3), that $|\Lambda^{-1}u_\varepsilon|$ and $|\Lambda^{1/2}u_\varepsilon|$ are bounded too, and then also $|u_\varepsilon|$.

We now return to (B1) to prove the key estimate: $|Du_\varepsilon|_\infty$ is bounded.

(B1) writes readily:

$$8\pi^2|\Lambda^{-1/2}u|^2+\int\frac{Du^2}{(1+Du^2)^{1/2}}\,dx\leqq A_5+2m\left(\varepsilon+(1+m^2)^{-1/2}\right)|u(0)|. \tag{4}$$

To proceed further we need a technical lemma.

LEMMA 3.5. *For every $K>1$, there is some $C(K)\geqq 0$ such that*

$$2|u(0)|\leqq K\int\frac{Du^2}{(1+Du^2)^{1/2}}\,dx+C(K)|u|^{2/3} \tag{5}$$

for every even function $u\in H^1(\mathbb{R})$.

Proof. Let $K>1$ be fixed, there is a $\lambda>0$ for which $x^2/(1+x^2)^{1/2}\geqq|x|/K$ whenever $|x|\geqq\lambda$. Then

$$\int_0^\infty\frac{Du^2}{(1+Du^2)^{1/2}}\,dx\geqq\frac{1}{K}\int_E|Du|\,dx+\frac{1}{\sqrt{1+\lambda^2}}\int_{E^c}|Du|^2\,dx$$

where $E=\{x\geqq 0:|Du(x)|\geqq\lambda\}$.

Suppose first that u is nonincreasing for $x\geqq 0$ and let

$$u_1(x)=-\int_x^\infty\chi_E Du\,dt,\qquad u_2(x)=-\int_x^\infty\chi_{E^c}Du\,dt,$$

where χ_E is the characteristic function of the set E. Obviously $u=u_1+u_2$ and $0\leqq u_2\leqq u$. As u_2^2 is absolutely continuous on every compact and $D(u_2^2)=2u_2Du_2$, we have:

$$u_2(0)^2=-2\int_0^\infty u_2Du_2\,dx\leqq 2\left(\int_0^\infty u^2\,dx\right)^{1/2}\left(\int_{E^c}Du^2\,dx\right)^{1/2}.$$

Applying Young's inequality, we get

$$u_2(0)\leqq\frac{K}{\sqrt{1+\lambda^2}}\int_{E^c}Du^2\,dx+C(K)\left(\int_0^\infty u^2\,dx\right)^{1/3}.$$

On the other hand

$$u_1(0)=-\int_E Du\,dx$$

and since u is even this proves (5).

The general case follows by considering the function $\tilde{u}(x)=\inf_{0\leqq y\leqq x}|u(y)|$ which is nonincreasing and satisfies $|\tilde{u}(0)|=|u(0)|$. $\tilde{u}$ is absolutely continuous on every compact set since $|\tilde{u}(x)-\tilde{u}(y)|\leqq|u(y)-u(z)|$ for some $z\in[x,y]$ for which $|u(z)|$ is minimum. As $\tilde{u}(x)\leqq|u(x)|$ and $|D\tilde{u}(x)|\leqq|Du(x)|$, applying (5) to $\tilde{u}$ (extended with $\tilde{u}(-x)=\tilde{u}(x)$) proves the lemma. □

Let us return to (4); for ε small enough we may apply Lemma 3.5 choosing K such that $(m(1+m^2)^{-1/2}+m\varepsilon)K=\beta<1$. We get

$$8\pi^2|\Lambda^{-1/2}u|^2+(1-\beta)\int\frac{Du^2}{(1+Du^2)^{1/2}}\,dx\leqq\text{constant},$$

and then

$$\int \frac{Du^2}{(1+Du^2)^{1/2}}\,dx \leqq \text{constant}.$$

On the other hand, since $Du \in H^1(0, +\infty)$, we have

$$D\log(1+Du^2) = 2\frac{DuD^2u}{1+Du^2};$$

thus for $x \geqq 0$

$$\log\bigl(1+Du(x)^2\bigr) - \log(1+m^2) = 2\int_0^x \frac{Du}{(1+Du^2)^{1/4}} \frac{D^2u}{(1+Du^2)^{3/4}}\,dt;$$

then

$$\bigl|\log(1+Du^2) - \log(1+m^2)\bigr|_\infty \leqq 2\left(\int \frac{Du^2}{(1+Du^2)^{1/2}}\,dx\right)^{1/2}\left(\int \frac{[D^2u]^2}{(1+Du^2)^{3/2}}\,dx\right)^{1/2}.$$

From the estimates above, this is bounded, consequently $|Du_\varepsilon|_\infty$ is bounded and so is $|[D^2u_\varepsilon]|$.

Let now ε go to zero. If $u_\varepsilon = u_0 + v_\varepsilon$, v_ε is bounded in H^2 and, passing to a subsequence if necessary, we may further assume that $v_\varepsilon \rightharpoonup v$ weakly in H^2.

Let $u = u_0 + v$. As $[D^2u_\varepsilon] \rightharpoonup [D^2u]$ weakly in L^2 and $Dv_\varepsilon \to Dv$ weakly in H^1 (and then also uniformly on every compact), $[D^2u_\varepsilon]/(1+Du_\varepsilon^2)^{3/2}$ must converge to $[D^2u]/(1+Du^2)^{3/2}$ weakly in L^2. Thus u satisfies equation (E), and since u_ε is even, so is u. To complete the proof of Theorem 3.1, it remains only to prove uniqueness.

Let u_1, u_2 be two solutions of (E) in the space $u_0 + H^2$; we have:

$$\Lambda^{-1}[u_1 - u_2 - x(Du_1 - Du_2)] - \frac{1}{4\pi^2}\left(\frac{[D^2u_1]}{(1+Du_1^2)^{3/2}} - \frac{[D^2u_2]}{(1+Du_2^2)^{3/2}}\right) = 0.$$

Multiplying by $u_1 - u_2$ and integrating by parts gives

$$2\bigl|\Lambda^{-1/2}(u_1 - u_2)\bigr|^2 + \frac{1}{4\pi^2}\int\left(\frac{Du_1}{(1+Du_1^2)^{1/2}} - \frac{Du_2}{(1+Du_2^2)^{1/2}}\right)(Du_1 - Du_2)\,dx = 0.$$

But $x/(1+x^2)^{1/2}$ is an increasing function of x and the integral is $\geqq 0$, so then $u_1 = u_2$.

Remark. The same mechanism working on an interface with no grain boundary and arbitrary initial profile leads us to solve the initial value problem:

$$\frac{\partial w}{\partial t} - \Lambda^1 \frac{D^2w}{(1+Dw^2)^{3/2}} = 0,$$
$$w(0,x) = w_0(x).$$

An existence uniqueness result for that initial value problem is given in [5].

REFERENCES

[1] R. R. COIFMAN, A. MCINTOSH AND Y. MEYER, *L'intégrale de Cauchy définit un opérateur borné sur L^2 pour les courbes lipschitziennes*, Ann. Math., 116 (1982), pp. 361–387.

[2] L. COUDURIER, N. EUSTATHOPOULOS, J. C. GJOUD AND P. DESRE, *Corrosion intergranulaire du cuivre par le plomb liquide sous l'effet des forces capillaires*, J. chimie physique, 74 (3) (1977), pp. 289–294.

[3] J. DUCHON AND R. ROBERT, *Solution de forme stationnaire pour une équation d'évolution non linéaire de la chimie des surfaces*, C. R. Acad. Sci. Paris, Sér. I, t. 292 (1981), pp. 547–549.

[4] _____, *Evolution d'une interface par capillarité et diffusion de volume I. Existence locale en temps*, Ann. Inst. Henri Poincaré, Analyse non linéaire Vol. 1, n°5, 1984, pp. 361–378.

[5] _____, *Evolution d'une interface solide-liquide sous l'effet de forces capillaires*, C. R. Acad. Sci. Paris, Série I, t. 293 (1981), pp. 183–185.

[6] W. W. MULLINS, *Grain boundary grooving by volume diffusion*, Transactions of the Metallurgical Society of AIME, 218 (1960), pp. 354–361.

[7] P. WITOMSKI, *Sur la résolution numérique de quelques problèmes non linéaires*, Thèse, Grenoble, France, 1983.

SIAM J. MATH. ANAL.
Vol. 17, No. 6, November 1986

012

RECONSTRUCTION FROM RESTRICTED RADON TRANSFORM DATA: RESOLUTION AND ILL-CONDITIONEDNESS*

W. R. MADYCH[†‡] AND S. A. NELSON[†]

Abstract. Estimates are obtained for the ill-conditionedness of the inversion problem, which are independent of the method of inversion. These estimates are in terms of the resolution required of the inversion algorithm and parameters that describe the restrictions. Two kinds of restrictions are considered: the limited angle type and the exterior (hole) type.

Key words. Radon transform, limited angle, exterior problem, resolution, reconstruction, ill-conditionedness

AMS(MOS) subject classification. Primary 44A15

1. Introduction. As demonstrated by the technology of X-ray tomography, Radon transform data can be inverted quite successfully in cases where a full range of data is available. However, success has been less notable in cases of restricted data such as exterior (hole) problems and limited angle problems. Here we consider certain examples which clarify some of the difficulties that occur when the data is so restricted. From these examples we obtain estimates which give lower bounds on the ill-conditionedness of the reconstruction problem for restricted data. These estimates involve a parameter b related to resolution and constants a, r related to the restrictions; they are independent of the method of reconstruction.

In the limited angle case, Davison [2] has used a singular value decomposition to study ill-conditionedness. At the end of §2 we make an effort to relate some of his results to ours. A singular value decomposition for the exterior problem has been carried out by Quinto [6]. Since the bibliographies in [2] and [6] provide an account of work related to inversion of restricted data, we only mention a few articles of an introductory nature: [7] and [8] for general tomography; [3] and [4] for restricted problems.

Our results extend to Radon transforms $\hat{f}$ of functions f on R^n but there are technical complications. For clarity, we do the case $n=2$ first and comment on generalizations in §3. If $n=2$ then $\hat{f}$ gives the result of integrating f along lines $\langle x,u\rangle = t$. Namely,

$$\hat{f}(u,t)=\int_{-\infty}^{\infty} f(tu+sv)\,ds.$$

Here $u=(u_1,u_2)$ is a point on the unit circle S^1 in R^2 and v is either of the two unit vectors perpendicular to u. For points $x=(x_1,x_2)$ and $y=(y_1,y_2)$ in the plane R^2, $\langle x,y\rangle = x_1y_1+x_2y_2$ denotes the usual inner (dot) product. For the exterior problem the

* Received by the editors December 21, 1983, and in revised form April 15, 1985. This research was partially supported by the National Science Foundation under grant MCS-8202147.
† Department of Mathematics, Iowa State University, Ames, Iowa 50011.
‡ Current address, Department of Mathematics, University of Connecticut, Storrs, Connecticut 06268.

data $\hat{f}$ is restricted to lines that miss a fixed disk. For the limited angle problem the vectors u are restricted to a fixed sector.

We let U denote the subset of $S^1 \times R$ that corresponds to these restrictions on (u,t). The distance from the point x_0 to the line $\langle x,u\rangle = t$ is $|t-\langle x_0,u\rangle|$. Thus

$$U=\{(u,t): |t-\langle x_0,u\rangle| > d\} \tag{1.1}$$

corresponds to the lines that miss the disk with center x_0 and radius d. For the limited angle case, we fix α in $(0,\pi/2)$ and take

$$U=\{(u,t): \langle u,e_1\rangle^2 \leqq \cos^2\alpha\} \tag{1.2}$$

where $e_1=(1,0)$. Note that $u=\pm(\cos\theta,\sin\theta)$, $\alpha \leqq \theta \leqq \pi-\alpha$ gives the vectors specified by (1.2).

The examples we consider are defined by

$$g_b(x)=\psi(x)e^{-ab|x|^2}\cos bx_1 \tag{1.3}$$

where $|x|^2=\langle x,x\rangle = x_1^2+x_2^2$ and a, b are positive numbers. We assume that ψ is a measurable function which is identically 1 on a disk $|x|<r$ and takes values in $[0,1]$ elsewhere. The function ψ is included in (1.3) to provide a way to make g_b vanish outside of a bounded set, but if desired, ψ may be taken to be identically 1.

Below, we assume $a>0$. However if one takes $a=0$ and takes ψ to be the indicator function for the unit disk then g_b reduces to a density that has been used in numerical tests of certain limited angle reconstruction algorithms; see [5].

We will be interested in g_b in cases where b becomes large with a and r fixed. In such cases the graph of g_b has furrows near the origin, running parallel to the x_2 axis. The spacing of the furrows is $2\pi/b$ and their length is of order $(ab)^{-1/2}$. One would expect g_b to be reasonably well reconstructed by algorithms claiming to resolve features whose size is of order $1/b$.

The following theorem shows that achieving such resolution when b is large, leads to severe demands on the accuracy of the data.

Indeed, this result gives an estimate on the accuracy required and, we stress, is independent of the inversion algorithm. Roughly speaking, it indicates that such data must be given with accuracy of order e^{-cb} in order to resolve features of size $1/b$; here c is a positive constant which depends on the geometry.

THEOREM 1.1. *Let g_b and r be as above. If U is defined by* (1.1) *with $x_0=(0,d)$, $d=\sqrt{2}\,r$ and if $a=(2\sqrt{2}\,r)^{-1}$ in* (1.3), *then for all $b>0$*

$$\sup_{(u,t)\in U} |\hat{g}_b(u,t)| \leqq 2\left(\frac{\pi}{ab}\right)^{1/2} e^{-ar^2b}. \tag{1.4}$$

Likewise, if U is defined by (1.2) *and if $a=(2r)^{-1}\sin\alpha$, then* (1.4) *again holds for all $b>0$.*

Proof. The final remark in §3 shows that for all (u,t)

$$|\hat{g}_b(u,t)| \leqq \left(\frac{\pi}{ab}\right)^{1/2}(e^{-abr^2}+e^{-abQ}) \tag{1.5}$$

where $Q=t^2+(u_2)^2(2a)^{-2}$. Thus it suffices to prove that $Q \geqq r^2$ for (u,t) in U. Since $|u|=1$, we see that $(u_1)^2 \leqq \cos^2\alpha$ implies $(u_2)^2 \geqq \sin^2\alpha$. Hence $Q \geqq (\sin\alpha)^2(2a)^{-2}=r^2$ in the case of U given by (1.2). In the other case we see from (1.1) that if (u,t) is in U then $|t-du_2| \geqq d$, which implies $t^2+(du_2)^2 \geqq d^2/2$. But this says $Q \geqq r^2$ because $d=\sqrt{2}\,r=(2a)^{-1}$.

The functions g_b can be used as building blocks to construct additional examples of densities g that satisfy (1.4). How this is done depends on whether U is given by (1.1) or (1.2).

If U is given by (1.2) then any translate of g_b will also satisfy (1.4). Superimposing such translates leads to examples like

$$g(x_1,x_2)=\int_{-1/2}^{1/2} g_b(x_1,x_2-t)\,dt. \tag{1.6}$$

For this example, the furrows are approximately of unit length when b is large. If U is given by (1.1) then other densities that satisfy (1.4) can be obtained from g_b by using rotations of the plane about the point $x_0=(0,d)$ in place of translations.

Further examples of densities that can be constructed by using the functions g_b are those of the form $f=h+cg_b$ where c is a constant and h is a suitable function. When b is large, (1.4) shows that there will only be slight differences in the U restricted Radon transform data for f and h. Since $|g_b(x)|\leqq\psi(x)e^{-ab|x|^2}$, c and h can be chosen so that f and h satisfy nonnegativity conditions or, more generally, conditions which specify a given range of values on a given subset.

2. Discussion. In this section we examine in more detail some of the implications of Theorem 1.1. We begin with some useful notation and estimates. Define $\phi_{a,b}$ on R^2 by

$$\phi_{a,b}(x_1,x_2)=\begin{cases} 1 & \text{if } \cos bx_1\geqq\tfrac{1}{2} \text{ and } ab|x|^2\leqq 1,\\ -1 & \text{if } \cos bx_1\leqq-\tfrac{1}{2} \text{ and } ab|x|^2\leqq 1,\\ 0 & \text{otherwise.}\end{cases}$$

The resemblance between $\phi_{a,b}$ and g_b is reflected in the fact that $\phi_{a,b}g_b\geqq 0$ and

$$|g_b|\geqq\frac{e^{-1}}{2}|\phi_{a,b}|\psi. \tag{2.1}$$

We will assume $abr^2\geqq 1$, so the factor of ψ in (2.1) can be dropped. The set where $|\phi_{a,b}(x)|=1$ consists of stripes inside a disk of radius $(ab)^{-1/2}$. Since the area of those stripes is at least half the area of the disk we have

$$\|\phi_{a,b}\|_p\geqq\left(\frac{\pi}{2ab}\right)^{1/p}. \tag{2.2}$$

Here $\|\ \|_p$ denotes the usual $L^p(R^2)$ norm:

$$\|f\|_p=\left\{\int|f(x)|^p\,dx\right\}^{1/p}$$

where $dx=dx_1dx_2$ is the usual measure on R^2 and the integral is taken over the whole plane.

Combining (2.1) and (2.2) gives the first inequality in

$$\frac{e^{-1}}{2}\left(\frac{\pi}{2ab}\right)^{1/p}\leqq\|g_b\|_p\leqq\left(\frac{\pi}{pab}\right)^{1/p}. \tag{2.3}$$

The second inequality follows from $|g_b(x)|\leqq e^{-ab|x|^2}$.

Now suppose V is a subset of U and A is a mapping of data $\hat{f}\,|_V$ into functions defined on some subset Ω of the plane. Here $\hat{f}\,|_V$ represents $\hat{f}$ restricted to the set V. A may be viewed as a reconstruction algorithm. For notational convenience, we denote the result of A applied to $\hat{f}\,|_V$ by Rf; symbolically $Rf = A(\hat{f}\,|_V)$. In addition, we define Rf to be 0 at points outside of Ω.

To see the implications of estimates like (1.4) let f_1 and f_2 be densities that, for some nonzero constant c, satisfy $f_1(x) - f_2(x) = cg_b(x)$ at all x in Ω and along all lines corresponding to V. If A is a reconstruction algorithm in the sense that Rf approximates f on Ω and if features of size $1/b$ are resolved, then on Ω one can expect $Rf_1 - Rf_2$ to resemble $c\phi_{a,b}$. This resemblance should be at least good enough to give an estimate such as

$$\| Rf_1 - Rf_2 \|_p \geq \frac{e^{-1}}{2} \left\{ \int_\Omega |c\phi_{a,b}(x)|^p \, dx \right\}^{1/p}.$$

This can be linked with an estimate like (2.2) if Ω is suitable. Even in the exterior case it is appropriate to assume that Ω contains the half disk D defined by $x_2 < 0$, $(x_1)^2 + (x_2)^2 < (ab)^{-1}$. Since the stripes where $|\phi_{a,b}(x)| = 1$ cover more than half the area of D, this assumption on Ω implies

$$\int_\Omega |\phi_{a,b}(x)|^p \, dx \geq \frac{\pi}{4ab}.$$

Combining the last two inequalities gives

$$\| Rf_1 - Rf_2 \|_p \geq \frac{|c|}{2e} \left(\frac{\pi}{4ab} \right)^{1/p}. \tag{2.4}$$

This difference in the reconstructions Rf_1, Rf_2 must be based on differences in the data $\hat{f}_1\,|_V$, $\hat{f}_2\,|_V$. By (1.4) the difference in the data is at most $2|c|(\pi/ab)^{1/2} e^{-ar^2 b}$. Thus to achieve (2.4), A would have to amplify such small differences between sets of data by a factor that is at least of order $(ab)^{1/2 - 1/p} e^{ar^2 b}$.

Such amplification means that, to keep the effects of noise down, the data will at least have to be accurate to roughly $O(e^{-ar^2 b})$. Note that $ar^2 = (r/2)\sin\alpha$ in the limited angle case and $ar^2 = (r/4)\sqrt{2}$ in the exterior case.

In the above discussion we have tried to minimize the assumptions made about A. For specific reconstruction algorithms A it should not be difficult to make the above considerations more precise.

For purposes of comparison with [2], we introduce some further notation. Let H_N denote the space of functions of the form $h = \chi\phi$ where χ is the indicator function of the unit disk in R^2 and ϕ is any polynomial on R^2 of degree no more than N. Let

$$\varepsilon_N(\alpha) = \min_{h \in H_N} \frac{\|\hat{h}\|_{2,\alpha}}{\|h\|_2} \tag{2.5}$$

where the denominator is the $L^2(R^2)$ norm of h and the numerator is defined by

$$\|\hat{h}\|_{2,\alpha}^2 = \frac{1}{2} \int_\alpha^{\pi - \alpha} \int_{-1}^{1} |\hat{h}(u(\theta), t)|^2 (1 - t^2)^{1/2} \, dt \, d\theta$$

with $u(\theta)=(\cos\theta,\sin\theta)$. The results in [2] imply that $\varepsilon_N(\alpha)=(\pi/N)\lambda_{N-1}(N,W)$ where $W=(\pi/2-\alpha)/\pi$ and $\lambda_{N-1}(N,W)$ is the function plotted in [2, Fig. 1]. As an analogue of $\varepsilon_N(\alpha)$ we consider

$$E_b(\alpha)=\frac{\sup|\hat{g}_b(u(\theta),t)|}{\|g_b\|_2}$$

where the sup is over $\alpha<\theta<\pi-\alpha$, $-\infty<t<\infty$ and where the parameter a in g_b is chosen so that $a=(2r)^{-1}\sin\alpha$. By (1.4) and (2.3) we have

$$E_b(\alpha)\leqq 4e\sqrt{2}\,e^{-(br\sin\alpha)/2}. \tag{2.6}$$

Figure 1 of [2] suggests that $\varepsilon_N(\alpha)$ varies with N and α in much the same way as the right side of (2.6) varies with b and α.

3. Estimates. In this section we establish the estimates that were used above and comment on extensions to higher dimensions. We use the natural n dimensional versions of the two-dimensional notation in §§1 and 2.

The k-plane transform $R_k(f)$ of a function f on R^n is defined by

$$R_k(f)(M)=\int_M f\,d\sigma$$

where M denotes any k-dimensional affine subspace of R^n, $1\leqq k\leqq n-1$ and $d\sigma$ is ordinary Euclidean measure on M. In what follows we let $q(M)$ denote the point on M that is closest to the origin. If $v_1,\cdots,v_k$ are an orthonormal basis for the subspace $M-q(M)$ then

$$R_k(f)(M)=\int\cdots\int f\left(q(M)+\sum_{i=1}^{k}s_iv_i\right)ds_1\cdots ds_k. \tag{3.1}$$

For $k=n-1$, this becomes the Radon transform and we use the notation $\hat{f}(u,t)=R_{n-1}(f)(M)$ where $M=\{x\in R^n\colon\langle x,u\rangle=t\}$.

PROPOSITION 3.1. *Let $h(x)=e^{-c|x|^2}\cos bx_1$ where b,c are positive constants and $x_1=\langle x,e_1\rangle$, $e_1=(1,0,\cdots,0)$. Then*

$$R_k(h)(M)=\left(\frac{\pi}{c}\right)^{k/2}e^{-b^2l^2(4c)^{-1}}e^{-c|q(M)|^2}\cos\left(b\langle q(M),e_1\rangle\right)$$

where $l^2=|P_Me_1|^2=\sum_{i=1}^{k}\langle v_i,e_1\rangle^2$ is the square of the length of the vector obtained by orthogonally projecting e_1 onto $M-q(M)$.

Remark. In the Radon transform case, $q(M)=tu$ and

$$|P_Me_1|^2=1-\langle u,e_1\rangle^2=\sum_{j=2}^{n}(u_j)^2.$$

Proof. If $x=q(M)+\sum_{i=1}^{k}s_iv_i$ with v_i as in (3.1), then $|x|^2=|q(M)|^2+\sum_{i=1}^{k}(s_i)^2$ and $x_1=\langle q(M),e_1\rangle+(s,a)$ where $(s,a)=\sum_{i=1}^{k}s_ia_i$ and $a_i=\langle v_i,e_1\rangle$. Thus, from (3.1)

$$R_k(h)(M)=\mathrm{Re}\int\cdots\int e^{-c|q(M)|^2}e^{-c|s|^2}e^{ib(\langle q(M),e_1\rangle+(s,a))}ds_1\cdots ds_k.$$

Applying $\int e^{-c|s|^2}e^{ib(s,a)}ds=(\pi/c)^{k/2}e^{-b^2|a|^2(4c)^{-1}}$ completes the proof.

PROPOSITION 3.2. *Let $g=\psi h$ where h is as in Proposition* 3.1 *and ψ is any measurable function on R^n taking values in the interval* $[0,1]$ *and satisfying $\psi(x)=1$ when $|x|<r$. Then*

$$|R_1(g-h)(M)|\leqq\left(\frac{\pi}{c}\right)^{1/2}e^{-cr^2} \tag{3.2}$$

while for $k\geqq 2$

$$|R_k(g-h)(M)|\leqq\frac{1}{2}\omega_{k-1}c^{-k/2}\int_{cr^2}^{\infty}e^{-u}u^{(k/2)-1}\,du \tag{3.3}$$

where ω_{k-1} denotes the area of the unit sphere in R^k.

Remark. The asymptotic expansion [1, pp. 13–14]

$$\int_z^{\infty}e^{-u}u^{q-1}\,du\sim z^q e^{-z}\left\{\frac{1}{z}+\frac{q-1}{z^2}+\frac{(q-1)(q-2)}{z^3}+\cdots\right\} \tag{3.4}$$

shows that (3.3) gives bounds much like (3.2) when cr^2 is large. If q is a positive integer (k even) then (3.4) is exact. In particular, if $k=2$ then (3.3) says

$$|R_2(g-h)(M)|\leqq\left(\frac{\pi}{c}\right)e^{-cr^2}. \tag{3.5}$$

Proof. Using the fact that $|g-h|(x)\leqq\chi(x)e^{-c|x|^2}$ where χ is the indicator function for the set $|x|\geqq r$ we see from (3.1) that

$$|R_k(g-h)(M)|\leqq J(k,c,r,|q(M)|) \tag{3.6}$$

where

$$J(k,c,r,a)=\int_{a^2+|s|^2\geq r^2}e^{-c(a^2+|s|^2)}\,ds_1\cdots ds_k. \tag{3.7}$$

We first show that for all real a

$$J(1,c,r,a)\leqq J(1,c,r,r) \tag{3.8}$$

and

$$J(k,c,r,a)\leqq J(k,c,r,0),\qquad k\geqq 2. \tag{3.9}$$

If $r^2\leqq a^2$ then the integration in (3.7) is over all of R^k and $e^{-c(a^2+s^2)}\leqq e^{-c(r^2+s^2)}$ so $J(k,c,r,a)\leqq J(k,c,r,r)$. Thus if (3.8) and (3.9) hold for $a^2\leqq r^2$ then they hold for all real a.

Assuming $r^2-a^2\geqq 0$ and changing to polar coordinates we have

$$J(k,c,r,a)=\omega_{k-1}\int_{(r^2-a^2)^{1/2}}^{\infty}e^{-c(a^2+\rho^2)}\rho^{k-1}\,d\rho$$

which becomes

$$J(k,c,r,a)=\frac{1}{2}\omega_{k-1}\int_{r^2}^{\infty}e^{-ct}(t-a^2)^{(k-2)/2}\,dt \tag{3.10}$$

after the change of variable $t=a^2+\rho^2$. If $t>r^2\geqq a^2$ then $(t-a^2)^{1/2}\leqq(t-r^2)^{1/2}$ and $(t-a^2)^{(k-2)/2}\leqq(t-0)^{(k-2)/2}$ for $k\geqq 2$. Thus (3.8) and (3.9) follow from (3.10).

To obtain (3.2) we apply (3.8) to (3.6) and note that $J(1,c,r,r)=(\pi/c)^{1/2}e^{-cr^2}$. To obtain (3.3) we apply (3.9) to (3.6) and express $J(k,c,r,0)$ using (3.10) and the change of variable $u=ct$.

Remark. To derive the inequality (1.5) write $|\hat{g}|\leqq|\hat{g}-\hat{h}|+|\hat{h}|$ and apply Propositions 3.1 and 3.2 with $c=ab$, $n=2$, $k=1$. A higher-dimensional version of (1.5) can be obtained in the same way but (3.3) gives it a messy appearance. If $n=3$, $k=2$ then (3.3) simplifies to (3.5), so that

$$|\hat{g}(u,t)|\leqq\frac{\pi}{c}\left(e^{-cr^2}+e^{-cQ}\right)$$

where $Q=b^2l^2(2c)^{-2}+t^2$. For $c=ab$ this gives the $n=3$ version of (1.5). If (1.1)–(1.3) are given $n=3$ interpretations, only two changes are needed in the statement of Theorem 1.1: take $x_0=de_0$ where e_0 is any unit vector orthogonal to e_1 and change the right side of (1.4) by replacing $(\pi/ab)^{1/2}$ with (π/ab). Concerning the proof we note that $|gt-d\langle e_0,u\rangle|\geqq d$ implies $t^2+d^2\langle e_0,u\rangle^2\geqq d^2/2$. Also, $\langle e_0,u\rangle^2+\langle e_1,u\rangle^2\leqq|u|^2=1$ so $\langle e_0,u\rangle^2\leqq 1-\langle e_1,u\rangle^2=l^2$. Thus, if (u,t) is in the set U given by (1.1) then $t^2+d^2l^2\geqq d^2/2$ and hence $Q\geqq r^2$. For the case of U given by (1.2), $Q\geqq r^2$ follows from $l^2\geqq\sin^2\alpha$.

Acknowledgment. We thank the referee for bringing reference [5] to our attention.

REFERENCES

[1] E. T. COPSON, *Asymptotic Expansions*, Cambridge Univ. Press, Cambridge, 1971.

[2] M. E. DAVISON, *The ill-conditioned nature of the limited angle tomography problem*, SIAM J. Appl. Math., 43 (1983), pp. 428–448.

[3] F. A. GRÜNBAUM, *The limited angle reconstruction problem*, in Computed Tomography, L. A. Shepp, ed., Proc. Symposia in Applied Mathematics, Vol. 27, American Mathematical Society, Providence, RI, 1983, pp. 43–61.

[4] R. M. LEWITT AND R. H. T. BATES, *Image reconstruction*: I, Optik, 50 (1978), pp. 19–33.

[5] A. PERES, *Tomographic reconstruction from limited angular data*, J. Comput. Assisted Tomography, 3 (1979), pp. 800–803.

[6] E. T. QUINTO, *Singular value decompositions and inversion methods for the exterior Radon transform and a spherical transform*, J. Math. Anal. Appl., 94 (1983), pp. 437–448.

[7] H. J. SCUDDER, *Introduction to computer aided tomography*, Proc. IEEE, 66 (1978), pp. 628–637.

[8] L. A. SHEPP AND J. B. KRUSKAL, *Computerized tomography: the new medical X-ray technology*, Amer. Math. Monthly, 85 (1978), pp. 420–439.

SIAM J. MATH. ANAL.
Vol. 17, No. 6, November 1986

013

SIGN VARIATIONS OF THE MACDONALD IDENTITIES*

DENNIS STANTON†

Abstract. Sign variations are given for the Macdonald identities for root systems of small rank. Limiting cases of these identities give properties of the eta function. Two such formulas are explicitly given.

Key words. Macdonald identities, eta function, root systems

AMS(MOS) subject classifications. Primary 10A20; secondary 10D05, 33A70

1. Introduction. The Macdonald identities [3] are the analogues of the Weyl denominator formula for affine root systems. Dyson [2], based upon Winquist's [6] original idea, found each infinite family of these identities (except $B_n^\vee$) by a case-by-case argument. Macdonald [3] gave a uniform proof of these identities.

Dyson's proof (and Macdonald's) can be thought of as having three main steps. Let $F(x_1,\cdots,x_n)$ be the product side of the Macdonald identities. Let F have the Laurent expansion

$$F(x_1,\cdots,x_n)=\sum_m f(m_1,\cdots,m_n;q)x_1^{m_1}\cdots x_n^{m_n}. \tag{1.1}$$

These three steps are:

(I) Use the affine part of the root system to give functional equations for $F(x_1,\cdots,x_n)$ or $f(m_1,\cdots,m_n;q)$. These equations reduce the number of unknown $f(m_1,\cdots,m_n;q)$ to a finite number. This gives a fundamental domain for the lattice M in the Macdonald identities (see [3]).

(II) Use the Weyl group W of the finite root system to show that only one function, $f(0,\cdots,0;q)$, the constant term, is unknown. The simple reflections give sign-reversing involutions which show that all terms other than the orbit of W on $(0,\cdots,0)$ are zero.

(III) Find the constant term by specializing the identity in (II). This can be done by choosing special roots of unity for each x_i and using the Weyl denominator and Jacobi triple product formulas.

In this paper we shall use this proof to give new identities, called sign variations. The infinite products in $F(x_1,\cdots,x_n)$ are modified by allowing appropriate signs. Step (I) still holds. Instead of the Weyl group in (II), a subgroup related to the signs is used. This time more than orbit survives. There is one constant to be computed for each orbit. The calculations can be carried out if the rank of the root system is small.

These steps also easily prove the generalization of the Macdonald identities of type A_n due to Milne [5]. They also indicate that the Macdonald identities could have a purely bijective proof.

* Received by the editors December 17, 1984, and in final revised form September 16, 1985. This research was partially supported by the National Science Foundation under grant MCS-8300872.

† School of Mathematics, University of Minnesota, Minneapolis, Minnesota 55455.

2. Type B_2. For Type B_2 the Macdonald identity is Winquist's identity [6], and the function $F(x,y)$ is

$$F(x,y)=(1/x)_\infty(qx)_\infty(1/y)_\infty(qy)_\infty(1/xy)_\infty(qxy)_\infty(y/x)_\infty(qx/y)_\infty, \tag{2.1}$$

where

$$(x)_\infty=(x;q)_\infty=\prod_{n=0}^{\infty}(1-xq^n). \tag{2.2}$$

(The base q will be omitted unless a base other than q is used.) The functional equations for (I) are

$$F(x,y)=-q^3x^3F(qx,y) \tag{2.3}$$

and

$$F(x,y)=-q^2y^3F(x,qy) \tag{2.4}$$

which imply that

$$f(i,j;q)=-q^if(i-3,j;q) \tag{2.5}$$

and

$$f(i,j;q)=-q^{j-1}f(i,j-3;q). \tag{2.6}$$

So the functions $f(i,j;q)$ are uniquely determined by the initial conditions at (i,j), $0\leqq i,\ j\leqq 2$.

For (II), the generators σ_1 and σ_2 of the Weyl group give

$$\sigma_1\colon\ -x^3F(x,y)=F(1/x,y) \tag{2.7}$$

and

$$\sigma_2\colon\ -xF(x,y)=yF(y,x). \tag{2.8}$$

These two equations imply

$$\sigma_1\colon\ f(i,j;q)=-f(-i-3,j;q) \tag{2.9}$$

and

$$\sigma_2\colon\ f(i,j;q)=-f(j-1,i+1;q). \tag{2.10}$$

Equations (2.9) and (2.10) give relations on the initial conditions. They can be recorded by Fig. 1: the edges of σ_1 and σ_2 are solid and dotted lines, respectively. The edges are labeled with the appropriate sign of the equality. It is clear that f is zero off the orbit of 00. Moreover, all terms in this orbit are determined by $f(0,0;q)$.

For (III), the function $F(x,y)$ is given by (II) as

(2.11)

$$F(x,y)=f(0,0;q)\sum_{i,j=-\infty}^{\infty}(-1)^{i+j}q^{(3i^2+3j^2+3i+j)/2}x^{3i}y^{3j}(1+x^2yq^{2i+j}+y^2q^{2j}-xyq^{i+j}).$$

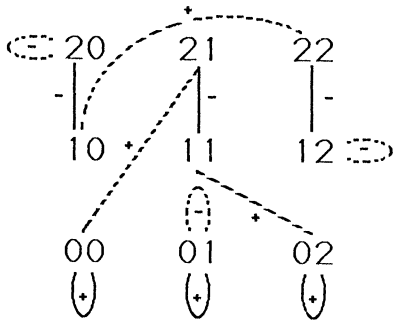


FIG. 1

We trisect (2.11) by replacing x and y independently by $-1, -\omega$, and $-\omega^2$, and summing the resulting equations. The Jacobi triple product identity then implies

$$f(0,0;q)=1/(q)_\infty^2. \tag{2.12}$$

For a sign variation, take a subset T of positive roots, and replace all infinite products $(z)_\infty$ in $F(x,y)$ involving T or $-T$ by $(-z)_\infty$. Then (I) still holds, with a possible change of sign. For (II), the subgroup $W(T)$ of W which fixes $T\cup -T$ is used. For example, if

$$F(x,y)=(1/x)_\infty(qx)_\infty(1/y)_\infty(qy)_\infty(1/xy)_\infty(qxy)_\infty(-y/x)_\infty(-qx/y)_\infty \tag{2.13}$$

then (2.3) and (2.4) hold with no minus sign. The involution σ_1 is replaced by

$$-x^2y^2F(x,y)=F(1/y,1/x), \tag{2.14}$$

and (2.8) has no minus sign. The graph that results is shown as Fig. 2. There are two orbits which survive: 00-11-02-21 and 20-12. The formula that results is

(2.15)

$$\begin{aligned} &F(x,y)\\ &=c_1(q)\sum_{i,j=-\infty}^{\infty} q^{(3i^2+3j^2+3i+j)/2}x^{3i}y^{3j}\left(1-q^{-j}/y+q^{-i-2j}/xy^2-q^{-2i-2j}/x^2y^2\right)\\ &\quad+c_2(q)\sum_{i,j=-\infty}^{\infty} q^{(3i^2+3j^2+3i+j)/2}x^{3i}y^{3j}\left(q^{-i}/x-q^{-2i-j}/x^2y\right), \end{aligned}$$

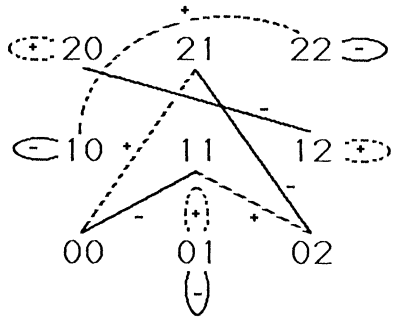


FIG. 2

where

$$c_1(q)=(q^3;q^3)_\infty(-q)_\infty/(-q^3;q^3)_\infty(q)_\infty^3 \tag{2.16}$$

and

$$c_2(q)=-2(-q^3;q^3)_\infty(q^6;q^6)_\infty/(q)_\infty^3. \tag{2.17}$$

There are two other B_2 sign variations. If

$$F(x,y)=(1/x)_\infty(qx)_\infty(1/y)_\infty(qy)_\infty(-1/xy)_\infty(-qxy)_\infty(y/x)_\infty(qx/y)_\infty, \tag{2.18}$$

then

(2.19)

$$F(x,y)=c_1(q)\sum_{i,j=-\infty}^{\infty}q^{(3i^2+3j^2+3i+j)/2}x^{3i}y^{3j}(1-q^{-j}/y-q^{-i-2j}/xy^2+q^{-2i-2j}/x^2y^2)$$

$$+c_2(q)\sum_{i,j=-\infty}^{\infty}q^{(3i^2+3j^2+3i+j)/2}x^{3i}y^{3j}(q^{-i-j}/xy-q^{-2i}/x^2),$$

where

$$c_1(q)=(q^3;q^3)_\infty(-q)_\infty/(-q^3;q^3)_\infty(q)_\infty^3 \tag{2.20}$$

and

$$c_2(q)=2(q^6;q^6)_\infty(-q^3;q^3)_\infty/(q)_\infty^3. \tag{2.21}$$

If

(2.22)

$$F(x,y)$$
$$=(-1/x)_\infty(-qx)_\infty(-1/y)_\infty(-qy)_\infty(-1/xy)_\infty(-qxy)_\infty(-y/x)_\infty(-qx/y)_\infty,$$

then

(2.23)

$$F(x,y)=c_1(q)\sum_{i,j=-\infty}^{\infty}q^{(3i^2+3j^2+3i+j)/2}x^{3i}y^{3j}(1+q^{-j}/y+q^{-i-2j}/xy^2+q^{-2i-2j}/x^2y^2)$$

$$+c_2(q)\sum_{i,j=-\infty}^{\infty}q^{(3i^2-3j^2+i+j)/2}x^{3i-1}y^{3j}(1+q^{-i}/x)(1+q^{-j}/y),$$

$$+c_3(q)\sum_{i,j=-\infty}^{\infty}q^{(3i^2+3j^2+3i-3j)/2}x^{3i}y^{3j-2},$$

where

$$c_1(q)=(q^3;q^3)_\infty^2(-q)_\infty^2/(-q^3;q^3)_\infty^2(q)_\infty^4, \tag{2.24}$$

$$c_2(q)=(q^3;q^3)_\infty(q^6;q^6)_\infty(-1)_\infty/(q)_\infty^4, \tag{2.25}$$

and

$$c_3(q)=2q^{1/2}\Big[(q^{3/2};q^3)_\infty(-q^{1/2})_\infty^2/(q^{1/2})_\infty - (-q^{3/2};q^3)_\infty(q^{1/2})_\infty^2/(-q^{1/2})_\infty\Big]/3(q)_\infty^2(q^3;q^6)_\infty^2. \tag{2.26}$$

In all of these formulas, on the sum side the orbit could be represented in another way. For example, in (2.11), the four terms could be taken as (x,y) with exponents $(3i,3j)$, $(3j-1,-3i-2)$, $(-3i-3,3j)$ and $(-3j-2,3i+1)$.

3. Types A_2, BC_1, and $B_2^\vee$. These are the remaining types which have at most three sign variations. Types G_2 and $G_2^\vee$ have 15 variations, while A_3 has 7. No infinite families of variations are given. It would appear that the appropriate constants are difficult to find in general.

Type A_2:

$$F(x,y)=(-1/x)_\infty(-qx)_\infty(1/y)_\infty(qy)_\infty(1/xy)_\infty(qxy)_\infty; \tag{3.1}$$

$$F(x,y)=c_1(q)\sum_{i,j=-\infty}^{\infty} q^{(i^2+j^2+ij+i+j)}(-1)^i(x^{2i+j}+x^{-i+j-1})y^{i+2j} + c_2(q)\sum_{i,j=-\infty}^{\infty} q^{(i^2+j^2+ij+3i+3j)}(-1)^i x^{2i+j+2}y^{i+2j+2} \tag{3.2}$$

where

$$c_1(q)=(q^3;q^3)_\infty(-q)_\infty/(-q^3;q^3)_\infty(q)_\infty^3 \tag{3.3}$$

and

$$c_2(q)=2q^3(-q^3;q^3)_\infty(q^6;q^6)_\infty/(q)_\infty^3. \tag{3.4}$$

Type BC_1:

$$F(x)=(-1/x)_\infty(-qx)_\infty(-qx^2;q^2)_\infty(-q/x^2;q^2)_\infty; \tag{3.5}$$

$$F(x)=c_1(q)\sum_{n=-\infty}^{\infty} q^{n(3n+1)/2}(x^{3n}+x^{-3n-1})+c_2(q)\sum_{n=-\infty}^{\infty} q^{3n(n+1)/2}x^{3n+1} \tag{3.6}$$

where

$$c_1(q)=(q^3;q^3)_\infty(-q^3;q^6)_\infty/(-q;q^2)_\infty(q)_\infty^2 \tag{3.7}$$

and

$$c_2(q)=2q(-q^6;q^6)_\infty(q^{12};q^{12})_\infty/(q^2;q^2)_\infty(q)_\infty. \tag{3.8}$$

Type $B_2^\vee$:

$$F(x,y)=(1/x^2;q^2)_\infty(q^2x^2;q^2)_\infty(1/y^2;q^2)_\infty(q^2y^2;q^2)_\infty(1/xy)_\infty(qxy)_\infty \times(-y/x)_\infty(-qx/y)_\infty; \tag{3.9}$$

$$F(x,y)=c_1(q)\sum_{i,j=-\infty}^{\infty} q^{2i^2+2i+2j^2+j}\left(x^{4i}y^{4j}-x^{-4j-3}y^{-4i-3}\right.$$
$$\left.-x^{-4i-4}y^{-4j-2}+x^{4j-1}y^{4i+1}\right)$$
$$+c_2(q)\sum_{i,j=-\infty}^{\infty} q^{2i^2+4i+2j^2+j}\left(x^{4i+2}y^{4j}-x^{-4j-3}y^{-4i-5}\right.$$
$$\left.-x^{-4i-6}y^{-4j-2}+x^{4j-1}y^{4i+3}\right) \tag{3.10}$$

where

$$c_1(q)=(-q^2;q^2)_\infty^5(-q)_\infty^3/(q^8;q^8)_\infty^2 \tag{3.11}$$

and

$$c_2(q)=-2q^2(-q^2;q^2)_\infty^3(-q)_\infty/(-q^2;q^4)_\infty^2(q)_\infty^2; \tag{3.12}$$

$$F(x,y)=(-1/x^2;q^2)_\infty(-q^2x^2;q^2)_\infty(1/y^2;q^2)_\infty(q^2y^2;q^2)_\infty$$
$$\times(1/xy)_\infty(qxy)_\infty(y/x)_\infty(qx/y)_\infty; \tag{3.13}$$

$$F(x,y)=c_1(q)\sum_{i,j=-\infty}^{\infty} q^{2i^2+2i+2j^2+j}(-1)^j x^{4i}\left(y^{4j}-y^{-4j-2}\right)$$
$$+c_2(q)\sum_{i,j=-\infty}^{\infty} q^{2i^2+3i+2j^2+2j}(-1)^j y^{4j+1}(x^{4i+1}+x^{-4i-5})$$
$$+c_3(q)\sum_{i,j=-\infty}^{\infty} q^{2i^2+4i+2j^2+j}(-1)^j x^{4i+2}\left(y^{4j}-y^{-4j-2}\right) \tag{3.14}$$

where

$$c_1(q)=(-q^2;q^2)_\infty^6/(q^8;q^8)_\infty^2(q;q^2)_\infty, \tag{3.15}$$

$$c_2(q)=-q(-q^2;q^2)_\infty\left[(q^{1/2})_\infty^4+(-q^{1/2})_\infty^4\right]/2(-q;q^2)_\infty^3(q^2;q^2)_\infty \tag{3.16}$$

and

$$c_3(q)=2q^2(-q^2;q^2)_\infty^6/(-q^2;q^4)_\infty^2(q^4;q^4)_\infty^2(q;q^2)_\infty; \tag{3.17}$$

$$F(x,y)=(1/x^2;q^2)(q^2x^2;q^2)_\infty(-1/y^2;q^2)_\infty(-q^2y^2;q^2)_\infty$$
$$\times(1/xy)_\infty(qxy)_\infty(y/x)_\infty(qx/y)_\infty; \tag{3.18}$$

$$F(x,y)=c_1(q)\sum_{i,j=-\infty}^{\infty} q^{2i^2+2i+2j^2+j}(-1)^i x^{4i}\left(y^{4j}+y^{-4j-2}\right)$$
$$+c_2(q)\sum_{i,j=-\infty}^{\infty} q^{2i^2+3i+2j^2+2j}(-1)^i y^{4j+1}(x^{4i+1}-x^{-4i-5})$$
$$+c_3(q)\sum_{i,j=-\infty}^{\infty} q^{2i^2+3i+2j^2+4j}(-1)^i y^{4j+3}(x^{4i+1}-x^{-4i-5}) \tag{3.19}$$

where

$$c_1(q)=\left[(q^{1/2})_\infty^4+(-q^{1/2})_\infty^4\right]/(q^2;q^2)_\infty^2(-q;q^2)_\infty^3, \tag{3.20}$$

$$c_2(q)=-q(-q^2;q^2)_\infty^6/(q^8;q^8)_\infty^2(q;q^2)_\infty, \tag{3.21}$$

and

$$c_3(q)=-2q^3(-q^2;q^2)_\infty^2(-q^4;q^4)_\infty^2/(q^2;q^2)_\infty(q)_\infty. \tag{3.22}$$

4. Remarks. One application of the Macdonald identities is to give expansions for powers of the eta function

$$\eta(q)=q^{1/24}(q)_\infty. \tag{4.1}$$

For example, if $x\to 1$ in the identity of type BC_1, then (see [3, p. 93])

$$\sum_{n\equiv 1 \pmod 6} nq^{n^2/24}=\eta(q)^5/\eta(q^2)^2. \tag{4.2}$$

If $x\to -1$ in the sign variation of BC_1 ((3.5) and (3.6)), we find

$$\sum_{n\equiv 1 \pmod 6} (-1)^{(n-1)/6}nq^{n^2/24}=\eta(q)\eta(q^2)^6\eta(q^{12})/\eta(q^4)^3\eta(q^6)^2 \tag{4.3}$$
$$+6\eta(q^2)\eta(q^3)^3\eta(q^{12})^3/\eta(q^4)\eta(q^6)^3.$$

As another example, we let $x,y\to 1$ in the sign variation (2.15) of B_2. It is easy to see that $\eta(q)^3c_1(q)$ and $\eta(q)^3c_2(q)$ can be expanded by the Jacobi triple product identity to obtain

(4.4)

$$4\eta(q)^7\eta(q^2)^2q^{-11/24}$$
$$=3\sum_{i,j,k=-\infty}^{\infty}\{i(3j-1)(3j+3i+1)+(3i+1)(3j+2)(i+j+1)\}q^{(3k^2+k+3i^2+3i+3j^2+j)/2}$$
$$-\sum_{i,j,k=-\infty}^{\infty}\{3j(3j-1)(3i-1)+(3i+1)(3i+2)(3j+2)\}q^{(3k^2+3k+3i^2+i+3j^2+j)/2}.$$

All of the sign variations can be treated in this way.

These techniques do not apply to the finite forms of the Macdonald conjectures in [4]. It is reasonable to ask if there are constant term formulas for finite forms of the sign variations. The answer is no, for in the low rank cases, such a result would imply that a well poised ${}_3F_2(-1)$ (see [1]) is evaluable.

REFERENCES

[1] R. ASKEY, *Some basic hypergeometric extensions of integrals of Selberg and Andrews*, this Journal, 11 (1980), pp. 938–951.

[2] F. DYSON, letter to L. Winquist, March, 1968.

[3] I. MACDONALD, *Affine root systems and Dedekind's η-function*, Inv. Math., 15 (1972), pp. 91–143.

[4] ———, *Some conjectures for root systems*, this Journal, 13 (1982), pp. 988–1007.

[5] S. MILNE, *An elementary proof of the Macdonald identities for* $A_1^{(1)}$, Adv. Math., 57 (1985), pp. 34–70.

[6] L. WINQUIST, *Elementary proof of* $p(11m+6)\equiv 0 \pmod{11}$, J. Comb. Theory A, 6(1969), pp. 56–59.

SIAM J. MATH. ANAL.
Vol. 17, No. 6, November 1986

014

PRODUCT AND ADDITION FORMULAS FOR THE CONTINUOUS q-ULTRASPHERICAL POLYNOMIALS*

MIZAN RAHMAN† AND ARUN VERMA‡

Abstract. Using Askey and Wilson's orthogonality relation and Rahman's product formula for Askey–Wilson polynomials a Gegenbauer-type product formula is obtained for continuous q-ultraspherical polynomials. Summation and transformation formulas for balanced hypergeometric series are then employed to derive a q-analogue of Gegenbauer's addition formula.

Key words. Rogers' q-ultraspherical polynomials, Askey–Wilson polynomials, product formula, addition formula, balanced and very-well-poised hypergeometric series, q-Saalschutz formula, Bailey's transformation formulas

AMS(MOS) subject classifications. Primary 33A65, 33A70; secondary 33A30

1. Introduction.

In 1875 Gegenbauer [9] gave the following addition formula for ultraspherical polynomials;

$$C_n^{\lambda}(\cos\psi)=\sum_{k=0}^{n} a_{k,n}^{\lambda}(\sin\theta)^k C_{n-k}^{\lambda+k}(\cos\theta)(\sin\phi)^k C_{n-k}^{\lambda+k}(\cos\phi) \cdot C_k^{\lambda-1/2}\left(\frac{\cos\psi-\cos\theta\cos\phi}{\sin\theta\sin\phi}\right), \tag{1.1}$$

where

$$a_{k,n}^{\lambda}=\frac{\Gamma(2\lambda-1)}{\Gamma^2(\lambda)}\,\frac{\Gamma^2(k+\lambda)(n-k)!(2k+2\lambda-1)2^{2k}}{\Gamma(n+k+2\lambda)} \tag{1.2}$$

and

$$C_n^{\lambda}(\cos\theta)=\sum_{k=0}^{n}\frac{(\lambda)_k(\lambda)_{n-k}}{k!(n-k)!}\cos(n-2k)\theta, \qquad 0\leqq\theta\leqq\pi. \tag{1.3}$$

See also Erdélyi [6, p. 178 (34), watch for the misprint 2^m which should be 2^{2m}], Askey [1, p. 30 (4.7), a factor of $n!/(2\lambda)_n$ is missing in the first term] or Whittaker and Watson [18, p.335, Ex. 42]. There are many proofs of this important formula, some are analytic in nature and some group theoretic. For extensive references see [1, Lecture 4]. One of the formulas that is contained in (1.1) is Gegenbauer's product formula

$$C_n^{\lambda}(\cos\theta)C_n^{\lambda}(\cos\phi)=b_n^{\lambda}\int_0^{\pi} C_n^{\lambda}(\cos\theta\cos\phi+\sin\theta\sin\phi\cos\psi)(\sin\psi)^{2\lambda-1}\,d\psi, \tag{1.4}$$

$\operatorname{Re}\lambda>0$, where

$$b_n^{\lambda}=C_n^{\lambda}(1)\left[\int_0^{\pi}(\sin\psi)^{2\lambda-1}\,d\psi\right]^{-1}=\frac{2^{1-2\lambda}\Gamma(2\lambda+n)}{\Gamma^2(\lambda)n!}, \tag{1.5}$$

* Received by the editors December 19, 1984, and in revised form August 26, 1985.

† Department of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada K1S 5B6. This work was supported by the Natural Sciences and Engineering Research Council of Canada under grant A6197.

‡ Department of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada K1S 5B6. Permanent address, Department of Mathematics, University of Roorkee, Roorkee, India.

see [5, p. 177, (20)] or [1, p. 30, (4.10)]. Equivalent to (1.4) is the product formula

$$\frac{C_n^\lambda(x)C_n^\lambda(y)}{C_n^\lambda(1)C_n^\lambda(1)}=\int_{-1}^{1}K(x,y,z)\frac{C_n^\lambda(z)}{C_n^\lambda(1)}\,dz, \tag{1.6}$$

where

$$K(x,y,z)=\frac{\Gamma(\lambda+1/2)}{\Gamma(\lambda)\Gamma(1/2)}\frac{(1-x^2-y^2-z^2+2xyz)^{\lambda-1}}{[(1-x^2)(1-y^2)]^{\lambda-1/2}}\quad\text{or } 0, \tag{1.7}$$

according as $1-x^2-y^2-z^2+2xyz$ is positive or negative, see [7, (1.4)] or [10].

The purpose of this paper is to obtain a q-analogue of (1.6) and then use it to derive an addition formula for Rogers' q-ultraspherical polynomials defined by

$$C_n(x;\beta\,|\,q)=\sum_{k=0}^{n}\frac{(\beta;q)_k(\beta;q)_{n-k}}{(q;q)_k(q;q)_{n-k}}\cos(n-2k)\theta,\qquad x=\cos\theta, \tag{1.8}$$

where the q-shifted factorials $(a;q)_k$ are given by

$$(a;q)_k=\frac{(a;q)_\infty}{(aq^k;q)_\infty},\qquad (a;q)_\infty=\prod_{j=0}^{\infty}(1-aq^j),\qquad |q|<1. \tag{1.9}$$

Askey and Ismail [2] have recently proved that

$$C_n(x;\beta\,|\,q)=\frac{(\beta;q)_n}{(q;q)_n}\beta^{-n/2}p_n(x;\sqrt{\beta},\sqrt{\beta q},-\sqrt{\beta},-\sqrt{\beta q}), \tag{1.10}$$

where

$$p_n(x;a,b,c,d)={}_4\phi_3\left[\begin{matrix}q^{-n},\,abcdq^{n-1},\,ae^{i\theta},\,ae^{-i\theta}\\ ab,\,ac,\,ad\end{matrix};q,q\right],\qquad x=\cos\theta \tag{1.11}$$

is the Askey–Wilson polynomial of degree n, discovered by Askey and Wilson [3], that satisfies the orthogonality relation

(1.12)

$$\int_{-1}^{1}w(x;a,b,c,d)p_m(x;a,b,c,d)p_n(x;a,b,c,d)\frac{dx}{\sqrt{1-x^2}}=h_n(a,b,c,d)\delta_{m,n},$$

with

$$w(x;a,b,c,d)=\frac{h(x;1)h(x;-1)h(x;\sqrt{q})h(x;-\sqrt{q})}{h(x;a)h(x;b)h(x;c)h(x;d)}, \tag{1.13}$$

$$h(x;a)=\prod_{n=0}^{\infty}(1-2axq^n+a^2q^{2n})=(ae^{i\theta};q)_\infty(ae^{-i\theta};q)_\infty, \tag{1.14}$$

$$h_n(a,b,c,d)=h_0(a,b,c,d)\frac{(q,cd,bd,bc;q)_n(1-abcdq^{-1})}{(abcdq^{-1},ab,ac,ad;q)_n(1-abcdq^{2n-1})}a^{2n}, \tag{1.15}$$

$$(a_1,a_2,\cdots,a_m;q)_n=\prod_{i=1}^{m}(a_i;q)_n \tag{1.16}$$

and

$$h_0(a,b,c,d)=\int_{-1}^{1} w(x;a,b,c,d)\frac{dx}{\sqrt{1-x^2}}=\frac{2\pi(abcd;q)_\infty}{(q,ab,ac,ad,bc,bd,cd;q)_\infty}, \tag{1.17}$$

provided $\max(|a|,|b|,|c|,|d|,|q|)<1$. The symbol on the right side of (1.11) is a special type of basic hypergeometric series defined by

$${}_{r+1}\phi_r\left[\begin{matrix}a_1,a_2,\cdots,a_{r+1}\\ b_1,\cdots,b_r\end{matrix};q,z\right]=\sum_{k=0}^{\infty}\frac{(a_1,a_2,\cdots,a_{r+1};q)_k}{(q,b_1,\cdots,b_r;q)_k}z^k, \tag{1.18}$$

whenever the series converges. The series becomes a polynomial of degree n in z if any one of the numerator parameters has the form q^{-n}, $n=0,1,2,\cdots$. The series (1.18) is called balanced if $z=q$ and $b_1b_2\cdots b_r=qa_1a_2\cdots a_{r+1}$; it is called well-poised if $a_2b_1=a_3b_2=\cdots=a_{r+1}b_r=qa_1$; and very-well-poised if, in addition, $b_1=\sqrt{a_1}$, $b_2=-\sqrt{a_1}$. Note that the ${}_4\phi_3$ series on the right of (1.11) is balanced.

In addition to balanced ${}_4\phi_3$ series we make fairly extensive use of very-well-poised series in this paper, so we shall adopt an economical notation:

$${}_{r+1}\phi_r\left[\begin{matrix}a,q\sqrt{a},-q\sqrt{a},a_1,a_2,\cdots,a_{r-2}\\ \sqrt{a},-\sqrt{a},aq/a_1,aq/a_2,\cdots,aq/a_{r-2}\end{matrix};q,z\right] \tag{1.19}$$

$$\equiv {}_{r+1}W_r(a;a_1,a_2,\cdots,a_{r-2};q,z).$$

In §2 we shall use a q-analogue of Koorwinder's proof [12] of his product formula for Jacobi polynomials to prove that a q-analogue of (1.6) is

$$p_n(x;a,a\sqrt{q},-a,-a\sqrt{q})\,p_n(y;a,a\sqrt{q},-a,-a\sqrt{q}) \tag{1.20}$$

$$=\int_{-1}^{1} K(x,y,z\,|\,q)\,p_n(z;a,a\sqrt{q},-a,-a\sqrt{q})\,dz,$$

where

$$K(x,y,z\,|\,q)=A^{-1}(\theta,\phi)(1-z^2)^{-1/2}w(z;ae^{i\theta+i\phi},ae^{-i\theta-i\phi},ae^{i\theta-i\phi},ae^{i\phi-i\theta}), \tag{1.21}$$

and

$$A(\theta,\phi)=h_0(ae^{i\theta+i\phi},ae^{-i\theta-i\phi},ae^{i\theta-i\phi},ae^{i\phi-i\theta}) \tag{1.22}$$

$$=\frac{2\pi(a^4;q)_\infty}{(q,a^2,a^2;q)_\infty|(a^2e^{2i\theta},a^2e^{2i\phi};q)_\infty|^2},$$

$x=\cos\theta$, $y=\cos\phi$, $0\leqq\theta\leqq\pi$, $0\leqq\phi\leqq\pi$, $\max(|a|,|q|)<1$. Note that $K(x,y,z\,|\,q)$ is positive for all $z\in(-1,1)$ and all θ,ϕ in $[0,\pi]$, unlike $K(x,y,z)$ in (1.7) which vanishes unless x,y,z satisfy certain triangle-type inequalities. However, after replacing a by q^a and then a by $\lambda/2$ we will show in §3 that

$$\lim_{q\to 1-} K(x,y,z\,|\,q)=K(x,y,z). \tag{1.23}$$

It may be pointed out that (1.20) is probably not optimal as a positive kernel for q-ultraspherical expansions since $p_n(x;q,a\sqrt{q},-a,-a\sqrt{q})$ does not equal 1 at $x=1$, although

$$\lim_{q\to 1-} p_n(x;q^{\lambda/2},q^{(\lambda+1)/2},-q^{\lambda/2},-q^{(\lambda+1)/2}) = C_n^{\lambda}(x)/C_n^{\lambda}(1).$$

In §4 we will make use of the orthogonality property (1.12) of Askey–Wilson polynomials to obtain the addition formula

(1.24)

$$\begin{aligned} &p_n(z;a,a\sqrt{q},-a,-a\sqrt{q}) \\ &= \sum_{m=0}^{n} C_{m,n}(a\mid q)e^{-mi\theta}(a^2e^{2i\theta};q)_m p_{n-m}(x;aq^{m/2},aq^{(m+1)/2},-aq^{m/2},-aq^{(m+1)/2}) \\ &\quad\cdot e^{-mi\phi}(a^2e^{2i\phi};q)_m p_{n-m}(y;aq^{m/2},aq^{(m+1)/2},-aq^{m/2},-aq^{(m+1)/2}) \\ &\quad\cdot p_m(z;ae^{i\theta+i\phi},ae^{-i\theta-i\phi},ae^{i\theta-i\phi},ae^{i\phi-i\theta}), \end{aligned}$$

where

$$C_{m,n}(a\mid q)=\frac{(q;q)_n(a^2,a^4q^n,a^4q^{-1};q)_m q^{m(m-n)}}{(q;q)_m(q;q)_{n-m}(a^2\sqrt{q},-a^2\sqrt{q},-a^2;q)_m(a^4q^{-1};q)_{2m}}. \tag{1.25}$$

Writing $x=\cos\theta$, $y=\cos\phi$, $z=\cos\psi$, replacing a by $q^{\lambda/2}$ and taking the limit $q\to 1-$ one can easily see that (1.24) goes directly to (1.1) after using the ${}_2F_1$ representation of $C_n^{\lambda}(x)$:

$$\frac{C_n^{\lambda}(x)}{C_n^{\lambda}(1)}={}_2F_1\left(-n,n+2\lambda;\lambda+\frac{1}{2};\frac{1-x}{2}\right). \tag{1.26}$$

In a subsequent paper we shall consider some special cases of (1.24) and discuss some applications.

2. Proof of the product formula (1.20). The starting point for the derivation of (1.20) is Rahman's product formula [8], [14] for Askey–Wilson polynomials which, in the q-ultraspherical case, can be written as

(2.1)

$$\begin{aligned} &p_n(x;a,a\sqrt{q},-a,-a\sqrt{q})p_n(y;a,a\sqrt{q},-a,-a\sqrt{q}) \\ &=\frac{1+a^2q^n}{1+a^2}\left(-\frac{1}{\sqrt{q}}\right)^n\sum_{k=0}^{n}\frac{(q^{-n},a^4q^n,-a\sqrt{q}e^{i\theta},-a\sqrt{q}e^{-i\theta},-a\sqrt{q}e^{i\phi},-a\sqrt{q}e^{-i\phi};q)_k}{(q,a^2\sqrt{q},-a^2\sqrt{q},-a^2\sqrt{q},-a^2q,-\sqrt{q};q)_k} \\ &\quad\cdot q^k{}_{10}W_9(-q^{-k-1/2};ae^{i\theta},ae^{-i\theta},ae^{i\phi},ae^{-i\phi},-q^{-k}/a^2,q^{1/2-k}/a^2,q^{-k};q,q). \end{aligned}$$

The important property of the very-well-poised series on the right side is that it is balanced and terminating, so we can apply Bailey's transformation formula [4, p. 68, 8.5 (1)] as often as necessary.

In particular, we have

(2.2)

$$
\begin{aligned}
&{}_{10}W_9\left(-q^{-k-1/2};ae^{i\theta},ae^{-i\theta},ae^{i\phi},ae^{-i\phi},-q^{-k}/a^2,q^{1/2-k}/a^2,q^{-k};q,q\right)\\
&\quad=\frac{\left(-\sqrt{q},-a^2\sqrt{q},-ae^{i\phi},-ae^{-i\phi};q\right)_k}{\left(-1,-a^2,-a\sqrt{q}\,e^{i\phi};-a\sqrt{q}\,e^{-i\phi};q\right)_k}\\
&\qquad\cdot{}_{10}W_9\left(-q^{-k};a\sqrt{q}\,e^{i\theta},a\sqrt{q}\,e^{-i\theta},ae^{i\phi},ae^{-i\phi},q^{1/2-k}/a^2,-q^{1/2-k}/a^2,q^{-k};q,q\right).
\end{aligned}
$$

Thus

(2.3)

$$
\begin{aligned}
&p_n\left(x;a,a\sqrt{q},-a,-a\sqrt{q}\right)p_n\left(y;a,a\sqrt{q},-a,-a\sqrt{q}\right)\\
&=\frac{1+a^2q^n}{1+a^2}(-1)^nq^{-n/2}\sum_{k=0}^{n}\frac{\left(q^{-n},a^4q^n,-a\sqrt{q}\,e^{i\theta},-a\sqrt{q}\,e^{-i\theta},-ae^{i\phi},-ae^{-i\phi};q\right)_k}{\left(q,a^2\sqrt{q},-a^2,-a^2\sqrt{q},-a^2q,-1;q\right)_k}q^k\\
&\quad\cdot{}_{10}W_9\left(-q^{-k};a\sqrt{q}\,e^{i\theta},a\sqrt{q}\,e^{-i\theta},ae^{i\phi},ae^{-i\phi},q^{1/2-k}/a^2,-q^{1/2-k}/a^2,q^{-k};q,q\right).
\end{aligned}
$$

The key step now is to use Bailey's transformation formula [17, p. 101, (3.4.1.6)] between a balanced nearly-poised ${}_5\phi_4$ series and a very-well-poised balanced ${}_{12}\phi_{11}$. In fact, we only need a special case of this formula giving a balanced nearly-poised ${}_4\phi_3$ series in terms of a balanced and very-well-poised ${}_{10}\phi_9$ series:

(2.4)

$$
\begin{aligned}
&{}_4\phi_3\left[\begin{matrix}A,B,C,q^{-k}\\Aq/B,Aq/C,B^2C^2q^{-k-1}/A\end{matrix};q,q\right]\\
&=\frac{\left(\frac{q\sqrt{Aq}}{BC},\frac{A^2q^2}{B^2C^2},\sqrt{Aq},-\frac{\sqrt{Aq}}{B},-\frac{\sqrt{Aq}}{C};q\right)_k}{\left(\frac{Aq^2}{B^2C^2},-\frac{Aq}{BC},\frac{Aq}{B},\frac{Aq}{C},-1;q\right)_k}\\
&\cdot{}_{10}W_9\left(-q^{-k};-\sqrt{Aq},-q\frac{\sqrt{Aq}}{BC},\frac{\sqrt{Aq}}{B},\frac{\sqrt{Aq}}{C},\frac{BCq^{-k-1/2}}{A},-\frac{BCq^{-k-1/2}}{A},q^{-k};q,q\right).
\end{aligned}
$$

It may be pointed out that the right side does not follow directly from Bailey's above-mentioned formula, but only after we make use of Bailey's other formula [4, p. 68, 8.5(1)].

Clearly, if we set $\sqrt{A}=-ae^{i\theta}$, $B=-\sqrt{q}\,e^{i\theta+i\phi}$, $C=-\sqrt{q}\,e^{i\theta-i\phi}$, we obtain

(2.5)
$$
\begin{aligned}
&{}_{10}W_9\left(-q^{-k};a\sqrt{q}\,e^{i\theta},a\sqrt{q}\,e^{-i\theta},ae^{i\phi},ae^{-i\phi},q^{1/2-k}/a^2,-q^{1/2-k}/a^2,q^{-k};q,q\right)\\
&\quad=\frac{\left(-1,-a^2,a^2e^{-2i\theta},-a^2\sqrt{q}\,e^{i\theta+i\phi},-a^2\sqrt{q}\,e^{i\theta-i\phi};q\right)_k}{\left(a^4,-a\sqrt{q}\,e^{i\theta},-a\sqrt{q}\,e^{-i\theta},-ae^{i\phi},-ae^{-i\phi};q\right)_k}\\
&\qquad\cdot{}_4\phi_3\left[\begin{matrix}q^{-k},a^2e^{2i\theta},-\sqrt{q}\,e^{i\theta+i\phi},-\sqrt{q}\,e^{i\theta-i\phi}\\q^{1-k}e^{2i\theta}/a^2,-a^2\sqrt{q}\,e^{i\theta-i\phi},-a^2\sqrt{q}\,e^{i\theta+i\phi}\end{matrix};q,q\right].
\end{aligned}
$$

For real a,θ,ϕ the expression on the left side of (2.5) is a real quantity, so the same must be true of the right-hand side which, however, is not self-evident. Since the ${}_4\phi_3$ series is balanced we can use an iteration of Sears' transformation formula [16]

$$(2.6)\quad {}_4\phi_3\left[\begin{matrix} q^{-k},a,b,c \\ d,e,q^{1-k}abc/de \end{matrix};q,q\right] = \frac{(e/a,de/bc;q)_k}{(e,de/abc;q)_k}\,{}_4\phi_3\left[\begin{matrix} q^{-k},a,d/b,d/c \\ d,de/bc,aq^{1-k}/e \end{matrix};q,q\right],\quad k=0,1,2,\cdots$$

to show that

$$(2.7)\quad {}_4\phi_3\left[\begin{matrix} q^{-k},a^2e^{2i\theta},-\sqrt{q}\,e^{i\theta+i\phi},-\sqrt{q}\,e^{i\theta-i\phi} \\ q^{1-k}e^{2i\theta}/a^2,-a^2\sqrt{q}\,e^{i\theta-i\phi},-a^2\sqrt{q}\,e^{i\theta+i\phi} \end{matrix};q,q\right] = \frac{(a^2,-a^2\sqrt{q}\,e^{i\phi-i\theta};q)_k}{(a^2e^{-2i\theta},-a^2\sqrt{q}\,e^{i\theta+i\phi};q)_k}\cdot{}_4\phi_3\left[\begin{matrix} q^{-k},a^2,-\sqrt{q}\,e^{i\theta+i\phi},-\sqrt{q}\,e^{-i\theta-i\phi} \\ q^{1-k}/a^2,-a^2\sqrt{q}\,e^{i\theta-i\phi},-a^2\sqrt{q}\,e^{i\phi-i\theta} \end{matrix};q,q\right].$$

Formulas (2.3), (2.5) and (2.7) then give

$$(2.8)\quad p_n(x;a,a\sqrt{q},-a,-a\sqrt{q})\,p_n(y;a,a\sqrt{q},-a,-a\sqrt{q}) = \frac{1+a^2q^n}{1+a^2}(-1)^nq^{-n/2}\sum_{k=0}^{n}\frac{(q^{-n},a^4q^n,a^2,-a^2\sqrt{q}\,e^{i\theta-i\phi},-a^2\sqrt{q}\,e^{i\phi-i\theta};q)_k}{(q,a^4,a^2\sqrt{q},-a^2\sqrt{q},-a^2q;q)_k}q^k\cdot{}_4\phi_3\left[\begin{matrix} q^{-k},a^2,-\sqrt{q}\,e^{i\theta+i\phi},-\sqrt{q}\,e^{-i\theta-i\phi} \\ q^{1-k}/a^2,-a^2\sqrt{q}\,e^{i\theta-i\phi},-a^2\sqrt{q}\,e^{i\phi-i\theta} \end{matrix};q,q\right].$$

Since by the q-Saalschutz formula [4, p. 68, 8.4(1)],

$$(2.9)\quad (-a\sqrt{q}\,e^{i\psi},-a\sqrt{q}\,e^{-i\psi};q)_k = (-a^2\sqrt{q}\,e^{i\theta-i\phi},-\sqrt{q}\,e^{i\phi-i\theta};q)_k\,{}_3\phi_2\left[\begin{matrix} q^{-k},ae^{i\theta-i\phi+i\psi},ae^{i\theta-i\phi-i\psi} \\ -a^2\sqrt{q}\,e^{i\theta-i\phi},-q^{1/2-k}e^{i\theta-i\phi} \end{matrix};q,q\right],$$

we find that, by (1.17) and (1.22)

$$(2.10)\quad \int_{-1}^{1} w(z;ae^{i\theta+i\phi},ae^{-i\theta-i\phi},ae^{i\theta-i\phi},ae^{i\phi-i\theta})(-a\sqrt{q}\,e^{i\psi},-a\sqrt{q}\,e^{-i\psi};q)_k\frac{dz}{\sqrt{1-z^2}} = (-a^2\sqrt{q}\,e^{i\theta-i\phi},-\sqrt{q}\,e^{i\phi-i\theta};q)_k\sum_{j=0}^{k}\frac{(q^{-k};q)_jq^j}{(q,-a^2\sqrt{q}\,e^{i\theta-i\phi},-q^{1/2-k}e^{i\theta-i\phi};q)_j}\cdot\int_{-1}^{1} w(z;ae^{i\theta+i\phi},ae^{-i\theta-i\phi},aq^je^{i\theta-i\phi},ae^{i\phi-i\theta})\frac{dz}{\sqrt{1-z^2}}$$

$$= A(\theta,\phi)\left(-a^2\sqrt{q}\,e^{i\theta-i\phi}, -\sqrt{q}\,e^{i\phi-i\theta}q\right)_k$$

$$\cdot {}_4\phi_3\left[\begin{matrix} q^{-k}, a^2, a^2e^{2i\theta}, a^2e^{-2i\phi} \\ -a^2\sqrt{q}\,e^{i\theta-i\phi}, a^4, -q^{1/2-k}e^{i\theta-i\phi}\end{matrix}; q, q\right]$$

$$= A(\theta,\phi)\frac{\left(a^2, -a^2\sqrt{q}\,e^{i\theta-i\phi}, -a^2\sqrt{q}\,e^{i\phi-i\theta}; q\right)_k}{\left(a^4; q\right)_k}$$

$$\cdot {}_4\phi_3\left[\begin{matrix} q^{-k}, a^2, -\sqrt{q}\,e^{i\theta+i\phi}, -\sqrt{q}\,e^{-i\theta-i\phi} \\ -a^2\sqrt{q}\,e^{i\theta-i\phi}, -a^2\sqrt{q}\,e^{i\phi-i\theta}, q^{1-k}/a^2\end{matrix}; q, q\right],$$

by (2.6). Combining (2.8) and (2.10) we then have

(2.11)

$$p_n(x; a, a\sqrt{q}, -a, -a\sqrt{q})\,p_n(y; a, a\sqrt{q}, -a, -a\sqrt{q})$$

$$= A^{-1}(\theta,\phi)\int_{-1}^{1}\frac{dz}{\sqrt{1-z^2}}w(z; ae^{i\theta+i\phi}, ae^{-i\theta-i\phi}, ae^{i\theta-i\phi}, ae^{i\phi-i\theta})$$

$$\cdot\frac{1+a^2q^n}{1+a^2}(-1)^n q^{-n/2}{}_4\phi_3\left[\begin{matrix} q^{-n}, a^4q^n, -a\sqrt{q}\,e^{i\psi}, -a\sqrt{q}\,e^{-i\psi} \\ a^2\sqrt{q}, -a^2\sqrt{q}, -a^2q\end{matrix}; q, q\right]$$

$$= A^{-1}(\theta,\phi)\int_{-1}^{1} w(z; ae^{i\theta+i\phi}, ae^{-i\theta-i\phi}, ae^{i\theta-i\phi}, ae^{i\phi-i\theta})$$

$$\cdot p_n(z; a, a\sqrt{q}, -a, -a\sqrt{q})\frac{dz}{\sqrt{1-z^2}} \quad \text{by (2.6).}$$

This completes the proof of (1.20). It follows from (1.20) that we have the expansion

(2.12)
$$\frac{\left(a^2; q\right)_\infty^3\left(qa^2; q\right)_\infty}{\left(a^4; q\right)_\infty^2}\left|\frac{\left(a^2e^{2i\theta}, a^2e^{2i\phi}, a^2e^{2i\psi}; q\right)_\infty}{\left(ae^{i\theta+i\phi+i\psi}, ae^{i\theta-i\phi+i\psi}, ae^{i\theta+i\phi-i\psi}, ae^{i\theta-i\phi-i\psi}; q\right)_\infty}\right|^2$$

$$\sim \sum_{n=0}^{\infty}\frac{\left(a^4; q\right)_n(1-a^2q^n)}{\left(q; q\right)_n(1-a^2)}a^{-2n}p_n(\cos\theta; a, a\sqrt{q}, -a, -a\sqrt{q})$$

$$\cdot p_n(\cos\phi; a, a\sqrt{q}, -a, -a\sqrt{q})\,p_n(\cos\psi; a, a\sqrt{q}, -a, -a\sqrt{q}).$$

By a direct computation it is shown by the authors in [15] that the " $\sim$ " sign in (2.12) can be replaced by " $=$ " for $\max(|a|, |q|) < 1$.

3. Limit of $K(x, y, z|q)$ as $q \to 1-$. Replacing a by q^a, $0 < q < 1$, in (1.21) and using the q-gamma function

(3.1)
$$\Gamma_q(x) = \frac{\left(q; q\right)_\infty}{\left(q^x; q\right)_\infty}(1-q)^{1-x}, \qquad \lim_{q\to 1-}\Gamma_q(x) = \Gamma(x),$$

it is not difficult to see that

$$\lim_{q\to 1-} K(x,y,z\,|\,q)$$

$$= \frac{\Gamma(4a)}{4\Gamma^2(2a)} \frac{\left|\sin\frac{\theta+\phi+\psi}{2}\sin\frac{\theta-\phi+\psi}{2}\sin\frac{\theta+\phi-\psi}{2}\sin\frac{\phi+\psi-\theta}{2}\right|^{2a-1}}{|\sin\theta\sin\phi|^{4a-1}}$$

$$\cdot \lim_{q\to 1-} L(\theta,\phi,\psi\,|\,q). \tag{3.2}$$

where

$$L(\theta,\phi,\psi\,|\,q)$$

$$= \frac{(\sqrt{q}\,;q)_\infty^2 |(\sqrt{q}\,e^{i\theta}, -\sqrt{q}\,e^{i\theta}, \sqrt{q}\,e^{i\phi}, -\sqrt{q}\,e^{-i\phi}, \sqrt{q}\,e^{i\psi}, -\sqrt{q}\,e^{-i\psi}; q)_\infty|^4}{|(\sqrt{q}\,e^{i(\theta+\phi+\psi)}, \sqrt{q}\,e^{i(\theta-\phi+\psi)}, \sqrt{q}\,e^{i(\theta+\phi-\psi)}, \sqrt{q}\,e^{i(\theta-\phi-\psi)}; q)_\infty|^2}. \tag{3.3}$$

Changing the base to q^2 and using the theta functions [18, Chap. 21] this can be written as

$$L(\theta,\phi,\psi\,|\,q^2)$$

$$= \left[\frac{(-q;q^2)_\infty}{(-q^2;q^2)_\infty}\right]^6 \frac{\vartheta_3^2\left(\frac{\theta}{2}\right)\vartheta_4^2\left(\frac{\theta}{2}\right)\vartheta_3^2\left(\frac{\phi}{2}\right)\vartheta_4^2\left(\frac{\phi}{2}\right)\vartheta_3^2\left(\frac{\psi}{2}\right)\vartheta_4^2\left(\frac{\psi}{2}\right)}{\vartheta_3^6\vartheta_4^2\vartheta_4\left(\frac{\theta+\phi+\psi}{2}\right)\vartheta_4\left(\frac{\theta-\phi+\psi}{2}\right)\vartheta_4\left(\frac{\phi+\psi-\theta}{2}\right)\vartheta_4\left(\frac{\theta+\phi-\psi}{2}\right)}, \tag{3.4}$$

where

$$\begin{aligned}\vartheta_3(\theta) &\equiv \vartheta_3(\theta\,|\,q) = (q^2;q^2)_\infty(-qe^{2i\theta};q^2)_\infty(-qe^{-2i\theta};q^2)_\infty,\\ \vartheta_4(\theta) &\equiv \vartheta_4(\theta\,|\,q) = (q^2;q^2)_\infty(qe^{2i\theta};q^2)_\infty(qe^{-2i\theta};q^2)_\infty\end{aligned} \tag{3.5}$$

and

$$\vartheta_3 = (-q;q^2)_\infty^2(q^2;q^2)_\infty, \qquad \vartheta_4 = (q,q)_\infty(q;q^2)_\infty. \tag{3.6}$$

Since

$$\lim_{q\to 1-} \frac{(-q;q^2)_\infty}{(-q^2;q^2)_\infty} = \sqrt{2} \tag{3.7}$$

we find that

$$\lim_{q\to 1-} K(x,y,z\,|\,q)$$

$$= \frac{2\Gamma(4a)}{\Gamma^2(2a)} \frac{\left|\sin\frac{\theta+\phi+\psi}{2}\sin\frac{\theta-\phi+\psi}{2}\sin\frac{\theta+\phi-\psi}{2}\sin\frac{\phi+\psi-\theta}{2}\right|^{2a-1}}{|\sin\theta\sin\phi|^{4a-1}}$$

$$\cdot \lim_{q\to 1-} M(\theta,\phi,\psi\,|\,q^2), \tag{3.8}$$

where

$$M(\theta,\phi,\psi\,|\,q^2) = \frac{\vartheta_3^2\left(\frac{\theta}{2}\right)\vartheta_4^2\left(\frac{\theta}{2}\right)\vartheta_3^2\left(\frac{\phi}{2}\right)\vartheta_4^2\left(\frac{\phi}{2}\right)\vartheta_3^2\left(\frac{\psi}{2}\right)\vartheta_4^2\left(\frac{\psi}{2}\right)}{\vartheta_3^6\vartheta_4^2\vartheta_4\left(\frac{\theta+\phi+\psi}{2}\right)\vartheta_4\left(\frac{\theta-\phi+\psi}{2}\right)\vartheta_4\left(\frac{\phi+\psi-\theta}{2}\right)\vartheta_4\left(\frac{\theta+\phi-\psi}{2}\right)}. \tag{3.9}$$

Using [18, p. 488, Ex. 2 and 3] we may write

$$M(\theta,\phi,\psi\,|\,q^2) = \left[1+H(\theta,\phi,\psi\,|\,q^2)\right]^{-1}, \tag{3.10}$$

where

$$H(\theta,\phi,\psi\,|\,q^2) = \sum_{i=1}^{11} A_i(\theta,\phi,\psi\,|\,q^2), \tag{3.11}$$

with

$$A_1 = \left[\frac{\vartheta_2\left(\frac{\theta}{2}\right)\vartheta_1\left(\frac{\psi}{2}\right)\vartheta_2\left(\frac{\psi}{2}\right)\vartheta_1\left(\frac{\theta}{2}\right)}{\vartheta_4\left(\frac{\theta}{2}\right)\vartheta_3\left(\frac{\psi}{2}\right)\vartheta_4\left(\frac{\psi}{2}\right)\vartheta_3\left(\frac{\theta}{2}\right)}\right]^2,$$

$$A_2 = 2\frac{\vartheta_2\left(\frac{\theta}{2}\right)\vartheta_1\left(\frac{\psi}{2}\right)\vartheta_2\left(\frac{\psi}{2}\right)\vartheta_1\left(\frac{\theta}{2}\right)}{\vartheta_4\left(\frac{\theta}{2}\right)\vartheta_3\left(\frac{\psi}{2}\right)\vartheta_4\left(\frac{\psi}{2}\right)\vartheta_3\left(\frac{\psi}{2}\right)},$$

$$A_3 = \left[\frac{\vartheta_4\vartheta_2\left(\frac{\theta}{2}\right)\vartheta_2\left(\frac{\psi}{2}\right)\vartheta_1^2\left(\frac{\phi}{2}\right)}{\vartheta_2\vartheta_4\left(\frac{\theta}{2}\right)\vartheta_4\left(\frac{\psi}{2}\right)\vartheta_3\left(\frac{\phi}{2}\right)}\right]^2,$$

$$A_4 = \left[\frac{\vartheta_4\vartheta_1\left(\frac{\theta}{2}\right)\vartheta_1\left(\frac{\phi}{2}\right)\vartheta_1\left(\frac{\psi}{2}\right)}{\vartheta_2\vartheta_3\left(\frac{\theta}{2}\right)\vartheta_3\left(\frac{\phi}{2}\right)\vartheta_3\left(\frac{\psi}{2}\right)}\right]^2, \tag{3.12}$$

$$A_5 = -2\left[\frac{\vartheta_4\vartheta_1\left(\frac{\phi}{2}\right)}{\vartheta_2\vartheta_3\left(\frac{\phi}{2}\right)}\right]^2\frac{\vartheta_2\left(\frac{\theta}{2}\right)\vartheta_1\left(\frac{\psi}{2}\right)\vartheta_2\left(\frac{\psi}{2}\right)\vartheta_1\left(\frac{\theta}{2}\right)}{\vartheta_4\left(\frac{\theta}{2}\right)\vartheta_3\left(\frac{\psi}{2}\right)\vartheta_4\left(\frac{\psi}{2}\right)\vartheta_3\left(\frac{\theta}{2}\right)},$$

$$A_6 = \left[\frac{\vartheta_4\vartheta_2\left(\frac{\psi}{2}\right)\vartheta_1\left(\frac{\theta}{2}\right)\vartheta_2\left(\frac{\phi}{2}\right)}{\vartheta_2\vartheta_4\left(\frac{\psi}{2}\right)\vartheta_3\left(\frac{\theta}{2}\right)\vartheta_4\left(\frac{\phi}{2}\right)}\right]^2,$$

$$A_7=\left[\frac{\vartheta_4\vartheta_2\left(\frac{\theta}{2}\right)\vartheta_1\left(\frac{\psi}{2}\right)\vartheta_2\left(\frac{\phi}{2}\right)}{\vartheta_2\vartheta_4\left(\frac{\theta}{2}\right)\vartheta_3\left(\frac{\psi}{2}\right)\vartheta_4\left(\frac{\phi}{2}\right)}\right]^2,$$

$$A_8=-2\left(\frac{\vartheta_4}{\vartheta_2}\right)^2\frac{\vartheta_2\left(\frac{\phi}{2}\right)\vartheta_2\left(\frac{\theta}{2}\right)\vartheta_1\left(\frac{\psi}{2}\right)\vartheta_2\left(\frac{\phi}{2}\right)\vartheta_1\left(\frac{\theta}{2}\right)\vartheta_2\left(\frac{\psi}{2}\right)}{\vartheta_4\left(\frac{\phi}{2}\right)\vartheta_4\left(\frac{\theta}{2}\right)\vartheta_3\left(\frac{\psi}{2}\right)\vartheta_4\left(\frac{\phi}{2}\right)\vartheta_3\left(\frac{\theta}{2}\right)\vartheta_4\left(\frac{\psi}{2}\right)},$$

$$A_9=\left[\frac{\vartheta_2\left(\frac{\phi}{2}\right)\vartheta_1\left(\frac{\theta}{2}\right)\vartheta_2\left(\frac{\theta}{2}\right)\vartheta_1\left(\frac{\phi}{2}\right)}{\vartheta_4\left(\frac{\phi}{2}\right)\vartheta_3\left(\frac{\theta}{2}\right)\vartheta_4\left(\frac{\theta}{2}\right)\vartheta_3\left(\frac{\phi}{2}\right)}\right]^2,$$

$$A_{10}=\left[\frac{\vartheta_2\left(\frac{\psi}{2}\right)\vartheta_1\left(\frac{\phi}{2}\right)\vartheta_2\left(\frac{\phi}{2}\right)\vartheta_1\left(\frac{\psi}{2}\right)}{\vartheta_4\left(\frac{\psi}{2}\right)\vartheta_3\left(\frac{\phi}{2}\right)\vartheta_4\left(\frac{\phi}{2}\right)\vartheta_3\left(\frac{\psi}{2}\right)}\right]^2,$$

$$A_{11}=-2\frac{\vartheta_2\left(\frac{\theta}{2}\right)\vartheta_1\left(\frac{\phi}{2}\right)\vartheta_2\left(\frac{\psi}{2}\right)\vartheta_1\left(\frac{\theta}{2}\right)\vartheta_2\left(\frac{\phi}{2}\right)\vartheta_1\left(\frac{\psi}{2}\right)\vartheta_2\left(\frac{\phi}{2}\right)\vartheta_1\left(\frac{\phi}{2}\right)}{\vartheta_4\left(\frac{\theta}{2}\right)\vartheta_3\left(\frac{\phi}{2}\right)\vartheta_4\left(\frac{\psi}{2}\right)\vartheta_3\left(\frac{\theta}{2}\right)\vartheta_4\left(\frac{\phi}{2}\right)\vartheta_3\left(\frac{\psi}{2}\right)\vartheta_4\left(\frac{\phi}{2}\right)\vartheta_3\left(\frac{\phi}{2}\right)}.$$

Using Poisson's transformation [18, p. 476]

$$\sum_{n=-\infty}^{\infty}e^{-n^2\pi t+2niz}=\frac{e^{-z^2/\pi t}}{\sqrt{t}}\sum_{n=-\infty}^{\infty}e^{-n^2\pi/t+2nz/t},\qquad \mathrm{Re}\,t>0 \tag{3.13}$$

and the definitions of the theta functions we find that, with $q=e^{-\pi t}$,

$$\begin{aligned}
\vartheta_1\left(\frac{\theta}{2}\right)&=\frac{e^{-(\pi-\theta)^2/4\pi t}}{\sqrt{t}}\sum_{n=-\infty}^{\infty}(-1)^n\exp\left[-\frac{\pi n^2}{t}-\frac{n(\pi-\theta)}{t}\right],\\
\vartheta_2\left(\frac{\theta}{2}\right)&=\frac{e^{-\theta^2/4\pi t}}{\sqrt{t}}\sum_{n=-\infty}^{\infty}(-1)^n\exp\left[-\frac{\pi n(n-\theta/\pi)}{t}\right],\\
\vartheta_3\left(\frac{\theta}{2}\right)&=\frac{e^{-\theta^2/4\pi t}}{\sqrt{t}}\sum_{n=-\infty}^{\infty}\exp\left[-\frac{\pi n(n-\theta/\pi)}{t}\right],\\
\vartheta_4\left(\frac{\theta}{2}\right)&=\frac{e^{-(\pi-\theta)^2/4\pi t}}{\sqrt{t}}\sum_{n=-\infty}^{\infty}\exp\left[-\frac{\pi n^2}{t}-\frac{n(\pi-\theta)}{t}\right],
\end{aligned} \tag{3.14}$$

see also [13, eq. (2.29)].

As $q \to 1-$ it follows that

$$
\begin{aligned}
&A_1 \simeq 1, \quad A_2 \simeq 2, \quad A_3 \simeq \left(\exp\left[\frac{\phi-\theta-\psi}{2t}\right]\right)^2, \\
&A_4 \simeq \left(\exp\left[\frac{\theta+\phi+\psi-2\pi}{t}\right]\right)^2, \qquad A_5 \simeq -2\left(\exp\left[\frac{\phi-\pi}{2t}\right]\right)^2, \\
&A_6 \simeq \left(\exp\left[\frac{\theta-\phi-\psi}{2t}\right]\right)^2, \qquad A_7 \simeq \left(\exp\left[\frac{\psi-\theta-\phi}{2t}\right]\right)^2, \\
&A_8 \simeq -2\exp\left(-\frac{\phi}{2t}\right), \quad A_9 \simeq 1, \quad A_{10} \simeq 1, \quad A_{11} \simeq -2.
\end{aligned} \tag{3.15}
$$

Thus

$$
\lim_{q \to 1-} H(\theta,\phi,\psi \mid q^2) = \begin{cases} \infty & \text{if either } \phi-\theta-\psi>0 \text{ or } \theta-\phi-\psi>0 \text{ or} \\ & \quad \psi-\theta-\phi>0 \text{ or } \theta+\phi+\psi-2\pi>0, \\ 3 & \text{if none of these is satisfied,} \end{cases} \tag{3.16}
$$

where $0 \leqq \theta, \phi, \psi \leqq \pi$.

Using (3.16) in (3.10) we find that

(3.17)

$$
\lim_{q \to 1} K(x,y,z \mid q)
$$

$$
= \begin{cases} 0 & \text{if either } \theta-\phi-\psi>0 \text{ or } \phi-\theta-\psi>0 \text{ or} \\ & \quad \psi-\theta-\phi>0 \text{ or } \theta+\phi+\psi-2\pi>0, \\ \dfrac{\Gamma(4a)}{2\Gamma^2(2a)} \dfrac{\left\{\sin\left(\frac{\theta+\phi-\psi}{2}\right)\sin\left(\frac{\phi+\psi-\theta}{2}\right)\sin\left(\frac{\psi+\theta-\phi}{2}\right)\sin\left(\frac{\theta+\phi+\psi}{2}\right)\right\}^{2a-1}}{(\sin\theta\sin\phi)^{4a-1}}, \end{cases}
$$

if none of the above inequalities is satisfied.

Setting $\lambda = 2a$ and using the duplication formula for the gamma function, it is easy to see that the right side of (3.17) reduces to that of (1.7).

4. Proof of the addition formula. Since by (1.12) the polynomials orthogonal with respect to the weight function $w(z; ae^{i\theta+i\phi}, ae^{-i\theta-i\phi}, ae^{i\theta-i\phi}, ae^{i\phi-i\theta})$ are $p_n(z; ae^{i\theta+i\phi}, ae^{-i\theta-i\phi}, ae^{i\theta-i\phi}, ae^{i\phi-i\theta})$, it is natural to look for an expansion of the form

$$
\begin{aligned}
&p_n(z; a, a\sqrt{q}, -a, -a\sqrt{q}) \\
&\quad = \sum_{m=0}^{n} A_{m,n}(\theta,\phi)\, p_m(z; ae^{i\theta+i\phi}, ae^{-i\theta-i\phi}, ae^{i\theta-i\phi}, ae^{i\phi-i\theta}).
\end{aligned} \tag{4.1}
$$

Using the orthogonality relation (1.12) we get

$$
\begin{aligned}
(4.2)\qquad & h_m(ae^{i\theta+i\phi}, ae^{-i\theta-i\phi}, ae^{i\theta-i\phi}, ae^{i\phi-i\theta})A_{m,n}(\theta,\phi) \\
&= \int_{-1}^{1} w(z; ae^{i\theta+i\phi}, ae^{-i\theta-i\phi}, ae^{i\theta-i\phi}, ae^{i\phi-i\theta}) \\
&\quad \cdot p_m(z; ae^{i\theta+i\phi}, ae^{-i\theta-i\phi}, ae^{i\theta-i\phi}, ae^{i\phi-i\theta}) \\
&\quad \cdot p_n(z; a, a\sqrt{q}, -a, -a\sqrt{q}) \frac{dz}{\sqrt{1-z^2}} \\
&= \frac{(1+q^n a^2)}{(1+a^2)}(-1)^n q^{-n/2} S_{m,n}(\theta,\phi),
\end{aligned}
$$

say, where, by (2.6),

$$
\begin{aligned}
(4.3)\qquad S_{m,n}(\theta,\phi) &= \int_{-1}^{1} w(z; ae^{i\theta+i\phi}, ae^{-i\theta-i\phi}, ae^{i\theta-i\phi}, ae^{i\phi-i\theta}) \\
&\quad \cdot p_m(z; ae^{i\theta+i\phi}, ae^{-i\theta-i\phi}, ae^{i\theta-i\phi}, ae^{i\phi-i\theta}) \\
&\quad \cdot p_n(z; -a\sqrt{q}, -a, a, a\sqrt{q}) \frac{dz}{\sqrt{1-z^2}}.
\end{aligned}
$$

By (1.11),

$$
(4.4)\qquad S_{m,n}(\theta,\phi) = \sum_{k=0}^{n}\sum_{j=0}^{m} \frac{(q^{-n}, a^4q^n; q)_k (q^{-m}, a^4q^{m-1}; q)_j q^{j+k} B_{j,k}}{(q, a^2\sqrt{q}, -a^2\sqrt{q}, -a^2q; q)_k (q, a^2, a^2e^{2i\theta}, a^2e^{2i\phi}; q)_j},
$$

where

$$
\begin{aligned}
(4.5)\qquad B_{j,k} &= \int_{-1}^{1} w(z; aq^j e^{i\theta+i\phi}, ae^{-i\theta-i\phi}, ae^{i\theta-i\phi}, ae^{i\phi-i\theta}) \\
&\quad \cdot (-a\sqrt{q}\, e^{i\psi}, -a\sqrt{q}\, e^{-i\psi}; q)_k \frac{dz}{\sqrt{1-z^2}}, \qquad z = \cos\psi.
\end{aligned}
$$

Except for the q^j factor in the first parameter in w, this integral is exactly the same as that in (2.10). Hence

$$
\begin{aligned}
(4.6)\qquad B_{j,k} &= A(\theta,\phi)(-a^2\sqrt{q}\, e^{i\theta-i\phi}, -\sqrt{q}\, e^{i\phi-i\theta}; q)_k (a^2, a^2e^{2i\theta}, a^2e^{2i\phi}; q)_j / (a^4; q)_j \\
&\quad \cdot \sum_{l=0}^{k} \frac{(q^{-k}, a^2, a^2e^{-2i\phi}, a^2q^j e^{2i\theta}; q)_l}{(q, -a^2\sqrt{q}\, e^{i\theta-i\phi}, -q^{1/2-k}e^{i\theta-i\phi}, a^4q^j; q)_l} q^l.
\end{aligned}
$$

Substituting in (4.4) we find that

$$
\begin{aligned}
(4.7)\qquad S_{m,n}(\theta,\phi) &= A(\theta,\phi) \sum_{k=0}^{n} \frac{(q^{-n}, a^4q^n, -a^2\sqrt{q}\, e^{i\theta-i\phi}, -\sqrt{q}\, e^{i\phi-i\theta}; q)_k}{(q, a^2\sqrt{q}, -a^2\sqrt{q}, -a^2q; q)_k} q^k \\
&\quad \cdot \sum_{l=0}^{k} \frac{(q^{-k}, a^2, a^2e^{2i\theta}, a^2e^{-2i\phi}; q)_l}{(q, a^4, -a^2\sqrt{q}\, e^{i\theta-i\phi}, -q^{1/2-k}e^{i\theta-i\phi}; q)_l} q^l \\
&\quad \cdot {}_3\phi_2\left[\begin{matrix} q^{-m}, a^4q^{m-1}, a^2e^{2i\theta}q^l \\ a^4q^l, a^2e^{2i\theta} \end{matrix}; q, q\right].
\end{aligned}
$$

By q-Saalschutz formula [4, p. 68, 8.4(1)], the ${}_3\phi_2$ series has the sum $(q^{1+l-m}, a^2e^{-2i\theta}; q)_m/(a^4q^l, q^{1-m}e^{-2i\theta}/a^2; q)_m$ which vanishes if $0 \leqq l \leqq m-1$. So the summation indices l and k in (4.7) both must be $\geqq m$. Replacing l and k by $l+m$ and $k+m$, respectively, and doing some straightforward simplification, we obtain

(4.8)

$$S_{m,n}(\theta,\phi) = A(\theta,\phi)\frac{\left(q^{-n}, a^4q^n, a^2, a^2e^{-2i\theta}, a^2e^{-2i\phi}; q\right)_m}{\left(a^2\sqrt{q}, -a^2\sqrt{q}, -a^2\sqrt{q}; q\right)_m (a^4; q)_{2m}} q^{m^2/2+m} a^{2m} e^{i(\theta+\phi)m}$$

$$\cdot \sum_{k=0}^{n-m} \frac{\left(q^{m-n}, a^4q^{n+m}, a^2q^m, -a^2q^{m+1/2}e^{i\theta-i\phi}, -a^2q^{m+1/2}e^{i\phi-i\theta}; q\right)_k}{\left(q, a^4q^{2m}, a^2q^{m+1/2}, -a^2q^{m+1/2}, -a^2q^{m+1}; q\right)_k} q^k$$

$$\cdot {}_4\phi_3\left[\begin{matrix} q^{-k}, a^2q^m, -\sqrt{q}\,e^{i\theta+i\phi}, -\sqrt{q}\,e^{-i\theta-i\phi} \\ q^{1-k-m}/a^2, -a^2q^{m+1/2}e^{i\theta-i\phi}, -a^2q^{m+1/2}e^{i\phi-i\theta} \end{matrix}; q, q\right].$$

Observe that the double series above is the same as that in (2.8) with n and a replaced by $n-m$ and $aq^{m/2}$, respectively. So we get

(4.9)

$$S_{m,n}(\theta,\phi) = A(\theta,\phi)\frac{\left(q^{-n}, a^4q^n, a^2, a^2e^{-2i\theta}, a^2e^{-2i\phi}; q\right)_m}{\left(a^2\sqrt{q}, -a^2\sqrt{q}, -a^2q; q\right)_m (a^4; q)_{2m}} q^{m^2/2+m} a^{2m} e^{i(\theta+\phi)m}$$

$$\cdot \frac{1+a^2q^m}{1+a^2q^n}(-1)^{n-m} q^{(n-m)/2} p_{n-m}\left(x; aq^{m/2}, aq^{(m+1)/2}, -aq^{m/2}, -aq^{(m+1)/2}\right)$$

$$\cdot p_{n-m}\left(y; aq^{m/2}, aq^{(m+1)/2}, -aq^{m/2}, -aq^{(m+1)/2}\right).$$

Using (1.15), (1.22) and (4.2), and simplifying, we finally obtain

$$A_{m,n}(\theta,\phi) = C_{m,n}(a\,|\,q)e^{-mi(\theta+\phi)}\left(a^2e^{2i\theta}, a^2e^{2i\phi}; q\right)_m \tag{4.10}$$

$$\cdot p_{n-m}\left(x; aq^{m/2}, aq^{(m+1)/2}, -aq^{m/2}, -aq^{(m+1)/2}\right)$$

$$\cdot p_{n-m}\left(y; aq^{m/2}, aq^{(m+1)/2}, -aq^{m/2}, -aq^{(m+1)/2}\right),$$

where $C_{m,n}(a\,|\,q)$ is given by (1.25). This completes the proof of the addition formula (1.24).

Acknowledgments. We would like to thank Professor G. Gasper and the referee for many valuable suggestions.

REFERENCES

[1] R. Askey, *Orthogonal Polynomials and Special Functions*, CBMS, Regional Conference Series in Applied Mathematics, Society for Industrial and Applied Mathematics, Philadelphia, 1975.

[2] R. Askey and M. E. H. Ismail, *A generalization of ultraspherical polynomials*, Studies in Pure Mathematics, P. Erdös, ed., Birkhauser, Boston, 1983, pp. 55–78.

[3] R. Askey and J. Wilson, *Some basic hypergeometric orthogonal polynomials that generalize Jacobi polynomials*, Mem. Amer. Math. Soc. 319, 1985.

[4] W. N. Bailey, *Generalized Hypergeometric Series*, Stechert-Hafner Service Agency, New York and London, 1964.

[5] A. ERDÉLYI, ed., *Higher Transcendental Functions*, Vol. I, McGraw-Hill, New York, 1953.
[6] _____, *Higher Transcendental Functions*, Vol. II, McGraw-Hill, New York, 1953.
[7] G. GASPER, *Positivity and the convolution structure for Jacobi series*, Ann. Math., 93 (1971), pp. 112–118.
[8] G. GASPER AND MIZAN RAHMAN, *Product formulas of Watson, Bailey and Bateman types and positivity of the Poisson kernel for q-Racah polynomials*, this Journal, 15 (1984), pp. 768–789.
[9] L. GEGENBAUER, *Über einige bestimmte Integrale*, Sitz. Math, Natur. Klasse, Akad. Wiss. Wien 70(2) (1975), pp. 433–443.
[10] H. Y. HSÜ, *Certain integrals and infinite sums involving ultraspherical polynomials and Bessel functions*, Duke Math. J., 4 (1938), pp. 374–383.
[11] M. E. H. ISMAIL AND J. A. WILSON, *Asymptotic and generating relations for the q-Jacobi and* ${}_4\phi_3$ *polynomials*, J. Approx. Theory, 36 (1982), pp. 43–54.
[12] T. H. KOORNWINDER, *Jacobi polynomials* II. *An analytic proof of the product formula*, this Journal, 5 (1974), pp. 125–137.
[13] B. NASSRALLAH AND MIZAN RAHMAN, *Projection formulas, a reproducing kernel and a generating function for q-Wilson polynomials*, this Journal, 16 (1985), pp. 186–197.
[14] MIZAN RAHMAN, *Reproducing kernels and bilinear sums for q-Racah and q-Wilson polynomials*, Trans. Amer. Math. Soc., 273 (1982), pp. 483–508.
[15] MIZAN RAHMAN AND ARUN VERMA, *Infinite sums of products of q-ultraspherical functions*, to appear.
[16] D. B. SEARS, *On the transformation theory of basic hypergeometric functions*, Proc. Lond. Math. Soc., 53 (1951), pp. 158–180.
[17] L. J. SLATER, *Generating Hypergeometric Functions*, Cambridge Univ. Press, Cambridge, 1966.
[18] E. T. WHITTAKER AND G. N. WATSON, *A Course of Modern Analysis*, 4th ed., Cambridge Univ. Press, Cambridge, 1927.

SIAM J. MATH. ANAL.
Vol. 17, No. 6, November 1986

015

ASYMPTOTICS OF THE ASKEY–WILSON AND q-JACOBI POLYNOMIALS*

MOURAD E. H. ISMAIL†

Abstract. We derive explicit representations and complete asymptotic expansions for the Askey–Wilson ${}_4\phi_3$ polynomials and the little and big q-Jacobi polynomials. We also give an alternate proof of a Dirichlet–Mehler type formula for the continuous q-ultraspherical polynomials. We also determine the asymptotic behavior of the q-Racah polynomials.

Key words. complete asymptotic expansion, little q-Jacobi polynomials, big q-Jacobi polynomials, Askey–Wilson polynomials, Dirichlet–Mehler formula

AMS(MOS) subject classifications. Primary 33A65, 41A60

1. Introduction.

The q-shifted factorial $(a;q)_n$ is defined by

$$(a;q)_0=1, \qquad (a;q)_n=\prod_{j=1}^{n}(1-aq^{j-1}), \qquad n=\infty,1,2,\cdots,|q|<1,$$

and the basic hypergeometric factorization ${}_r\phi_{r+p-1}$ is

$${}_r\phi_{r+p-1}\left(\begin{matrix}a_1,\cdots,a_r\\ b_1,\cdots,b_{r+p-1}\end{matrix};q,x\right)=\sum_{n=0}^{\infty}\frac{(a_1;q)_n\cdots(a_r;q)_n}{(b_1;q)_n\cdots(b_{r+p-1};q)_n}\frac{[(-1)^p x]^n}{(q;q)_n}q^{p\binom{n}{2}}.$$

The Askey–Wilson ${}_4\phi_3$ polynomials [6] are

$$p_n(x)=p_n(x;a,b,c,d):={}_4\phi_3\left(\begin{matrix}q^{-n},abcdq^{n-1},az,a/z\\ ab,ac,ad\end{matrix};q,q\right), \tag{1.1}$$

where

$$z^2-2xz+1=0, \qquad |z|\leqq|z^{-1}|; \tag{1.2}$$

that is

$$z=x-\sqrt{x^2-1}, \qquad \frac{1}{z}=x+\sqrt{x^2-1}. \tag{1.3}$$

In fact, if $x=\cos\theta$, then $z=e^{-i\theta}$. The big q-Jacobi polynomials of Andrews and Askey [2], [3] are

$$P_n(x)=P_n(x;\alpha,\beta,\gamma:q):={}_3\phi_2\left(\begin{matrix}q^{-n},\alpha\beta q^{n+1},x\\ \alpha q,\gamma q\end{matrix};q,q\right), \tag{1.4}$$

* Received by the editors February 20, 1985, and in revised form August 7, 1985. This research was partially supported by the National Science Foundation under grant MCS 8313931.

† Department of Mathematics, Arizona State University, Tempe, Arizona 85287. This work was done while the author was visiting the University of Minnesota, Minneapolis, Minnesota whose hospitality and support are gratefully acknowledged.

and the little q-Jacobi polynomials are defined via

$$\Phi_n^{\alpha,\beta}(x)={}_2\phi_1\left(\begin{matrix}q^{-n},\alpha\beta q^{n+1}\\ \alpha q\end{matrix};q,qx\right). \tag{1.5}$$

Ismail and Wilson [8] determined the main term in the asymptotic expansion of the aforementioned orthogonal polynomials for fixed x and $n\to\infty$. In this paper we determine the complete asymptotic expansions of these polynomials as $n\to\infty$. The idea is to apply Cauchy's theorem to a generating function and derive a contour integral representation, and then use analytic continuation and calculus of residues to evaluate the integral. This gives rise to a Dirichlet–Mehler type formula [12] which is also a convergent asymptotic series. In §2 we investigate the asymptotics of the Askey–Wilson ${}_4\phi_3$ polynomials. We also determine the main term in the asymptotic expansion, as $N\to\infty$, of $p_n(x;a,b,c,d)$ when $ab=q^{-N}$ and $n=O(N^{1-\varepsilon})$, where ε is a fixed number in $(0,1)$. The case $q=1$ of this later result is treated heuristically in [9]. The asymptotics of the little and big q-Jacobi polynomials are in §3. In §4 we give an alternate proof of a result of Rahman and Verma [10].

The Askey–Wilson ${}_4\phi_3$ polynomials generalize the $6-j$ symbols in the sense that they contain one more parameter and the $6-j$ symbols are limiting cases of ${}_4\phi_3$ polynomials. The Askey–Wilson polynomials are also the most general orthogonal polynomials that resemble the classical polynomials of Jacobi, Hermite and Laguerre. In fact, Andrews and Askey [2] define an orthogonal family of polynomials $\{r_n(x)\}$ to be classical if and only if $r_n(x)$ is a special case or a limiting case of the ${}_4\phi_3$ orthogonal polynomials.

2. The Askey–Wilson polynomials. Ismail and Wilson [8] derived the generating function

$$\sum_{n=0}^{\infty}\frac{t^n a^{-n}(ac,q)_n(ad;q)_n}{(cd;q)_n(q;q)_n}p_n(x)={}_2\phi_1\left(\begin{matrix}a/z,b/z\\ ab\end{matrix};q,zt\right){}_2\phi_1\left(\begin{matrix}cz,dz\\ cd\end{matrix};q,t/z\right). \tag{2.1}$$

The ${}_2\phi_1$'s in (2.1) converge when $|t|<|z|$ since $|z|\leqq|z^{-1}|$. We need to continue the ${}_2\phi_1$'s in (2.1) to meromorphic functions. The appropriate transformation to apply is the q-analogue of the Pfaff–Kummer transformation, [1],

$${}_2\phi_1\left(\begin{matrix}A,B\\ C\end{matrix};q,w\right)=\frac{(Aw;q)_\infty}{(w;q)_\infty}\sum_{k=0}^{\infty}\frac{(A;q)_k(C/B;q)_k q^{k(k-1)/2}}{(q;q)_k(C;q)_k(Aw;q)_k}(-Bw)^k. \tag{2.2}$$

This transforms the right-hand side of (2.1) to

$$\frac{(at;q)_\infty(ct;q)_\infty}{(zt;q)_\infty(t/z;q)_\infty}\sum_{k,j=0}^{\infty}\frac{(az;q)_k(a/z;q)_k(cz;q)_j(c/z;q)_j}{(q;q)_k(ab;q)_k(at;q)_k(q;q)_j(cd;q)_j}\cdot\frac{(-bt)^k(-dt)^j}{(ct;q)_j}q^{\binom{k}{2}+\binom{j}{2}}.$$

The above expression is analytic in the complex plane except at the poles $t=zq^{-m}$, $t=q^{-m}/z$, $m=0,1,\cdots$. Let us denote the left-hand side of (2.1) by $F(x,t)$. This gives the integral representation

$$\frac{(ac;q)_n(ad;q)_n}{(cd;q)_n(q;q)_n}p_n(x)=\frac{a^n}{2\pi i}\int_C t^{-n-1}F(x,t)\,dt, \tag{2.3}$$

where C is a circle $|t|=p$, $p<|z|$. Now think of C as a contour around the point $t=\infty$. So C encloses all the poles of $F(x,t)$ but has the wrong orientation. Therefore, the right-hand side of (2.3) is $-\sum$ Residues. Now $t=0$ is outside the contour. Observe that

$$\operatorname{Res}\{F(x,t)/t^{n+1};t=zq^{-m}\}=-\frac{(azq^{-m};q)_\infty(czq^{-m};q)_\infty q^{nm}}{(z^2q^{-m};q)_\infty(q^{-m};q)_m(q;q)_\infty}z^{-n}$$

$$\cdot\sum_{k,j=0}^{\infty}\frac{(az;q)_k(a/z;q)_k(cz;q)_j(c/z;q)_j b^k d^j(-zq^{-m})^{k+j}}{(q;q)_k(ab;q)_k(azq^{-m};q)_k(cd;q)_j(czq^{-m};q)_j(q;q)_j}$$

$$\cdot q^{j(j-1)/2+k(k-1)/2}.$$

This gives the asymptotic formula

(2.4)

$$\frac{(ac;q)_n(ad;q)_n}{(cd;q)_n(q;q)_n}p_n(x)$$

$$=\left(\frac{a}{z}\right)^n\frac{(az;q)_\infty(cz;q)_\infty}{(z^2;q)_\infty(q;q)_\infty}\sum_{m=0}^{\infty}\frac{\left(\frac{q}{az};q\right)_m\left(\frac{q}{cz};q\right)_m}{(q;q)_m(qz^{-2};q)_m}(ac)^m q^{mn}$$

$$\cdot{}_2\phi_2\left(\begin{matrix}a/z,az\\ab,azq^{-m}\end{matrix};q,zbq^{-m}\right){}_2\phi_2\left(\begin{matrix}cz,c/z\\cd,czq^{-m}\end{matrix};q,zdq^{-m}\right)$$

$$+\text{a similar term with } z \text{ and } 1/z \text{ interchanged.}$$

The relationship (2.4) is actually an explicit representation as well as an asymptotic expansion. As $n\to\infty$ the main term on the right-hand side of (2.4) is

$$(a/z)^n\frac{(az;q)_\infty(cz;q)_\infty}{(z^2;q)_\infty(q;q)_\infty}{}_1\phi_1\left(\begin{matrix}a/z\\ab\end{matrix};q,zb\right){}_1\phi_1\left(\begin{matrix}c/z\\cd\end{matrix};q,zd\right)$$

$$+\text{a similar term with } z \text{ and } 1/z \text{ interchanged.}$$

Clearly, if r is a positive integer, then

$${}_1\phi_1\left(\begin{matrix}a/z\\ab\end{matrix};q,zb\right)=\lim_{r\to\infty}{}_2\phi_1\left(\begin{matrix}a/z,q^{-r}\\ab\end{matrix};q,zbq^r\right)=\lim_{r\to\infty}\frac{(bz;q)_r}{(ab;q)_r}=\frac{(bz;q)_\infty}{(ab;q)_\infty},$$

where we used the q-analogue of Gauss's theorem, Slater [11, p. 247]. This shows that the main term in the asymptotic expansion of the left-hand side of (2.4) is

$$\left(\frac{a}{z}\right)^n A(z)+(az)^n A(z^{-1}),$$

where

$$A(z)=\frac{(az;q)_\infty(bz;q)_\infty(cz;q)_\infty(dz;q)_\infty}{(ab;q)_\infty(cd;q)_\infty(z^2;q)_\infty(q;q)_\infty}.$$

We now consider the q-Racah case, that is the case $ab=q^{-N}$. Formula (3.2) in [8] implies

$$\frac{(ac;q)_n(ad;q)_n}{(cd;q)_n(q;q)_n}p_n(x;a,q^{-N}/a,c,d)$$

$$=a^n\sum_{k=0}^{n}\frac{(a/z;q)_{n-k}(q^{-N}/az;q)_{n-k}(cz;q)_k(dz;q)_k}{(q^{-N};q)_{n-k}(q;q)_{n-k}(cd;q)_k(q;q)_k}z^{n-2k},$$

$0\leqq n\leqq N$. Applying

$$\frac{(q^{-N}/\lambda;q)_{n-k}}{(q^{-N};q)_{n-k}}=\frac{(\lambda q;q)_N(q;q)_{N-n+k}}{(\lambda q;q)_{N-n+k}(q;q)_N}\lambda^{k-n},$$

we obtain, $0\leqq n\leqq N$,

$$\text{(2.5)}\qquad \frac{(ac;q)_n(ad;q)_n}{(cd;q)_n(q;q)_n}p_n(x;a,q^{-N}/a,c,d)$$

$$=\frac{(aqz;q)_N}{(q;q)_N}\sum_{k=0}^{n}\frac{(cz;q)_k(dz;q)_k(q;q)_{N-n+k}(a/z;q)_{n-k}}{(cd;q)_k(q;q)_k(aqz;q)_{N-n+k}(q;q)_{n-k}}\left(\frac{a}{z}\right)^k.$$

We now let $N\to\infty$, $n\to\infty$ in such a way that $n=O(N^{1-\varepsilon})$, ε is a fixed number in $(0,1)$. One can easily justify interchanging the limiting and summation processes in (2.5) and establish

$$\text{(2.6)}\qquad p_n(x;a,q^{-N}/a,c,d)\cong\frac{(a/z;q)_\infty(cd;q)_\infty}{(ac;q)_\infty(ad;q)_\infty}{}_2\phi_1\left(\begin{matrix}cz,dz\\ad\end{matrix};q,a/z\right).$$

One disadvantage of the right-hand side in (2.6) is its lack of symmetry. It should be symmetric in z and $1/z$. In order to obtain a more symmetric representation we apply the transformation (2.2) and get

$$\text{(2.7)}\qquad p_n(x;a,q^{-N}/a,c,d)\cong\frac{(cd;q)_\infty}{(ad;q)_\infty}\sum_{k=0}^{\infty}\frac{(cz;q)_k(c/z;q)_k(-ad)^k}{(q;q)_k(cd;q)_k(ac;q)_k}q^{\binom{k}{2}},$$

as $n\to\infty$, $N\to\infty$, $n=O(N^{1-\varepsilon})$. Observe that $(cz;q)_k(c/z;q)_k$ is a polynomial in x of degree k. In fact

$$(cz;q)_k(c/z;q)_k=\prod_{j=0}^{k-1}(1-2q^jcx+c^2q^{zj}).$$

3. The q-Jacobi polynomials. Ismail and Wilson [8] found the generating function

$$\sum_{n=0}^{\infty}\frac{q^{n(n-1)/2}t^n}{(\beta q;q)_n(q;q)_n}\Phi_n^{\alpha,\beta}(x)={}_2\phi_1\left(\begin{matrix}0,1/x\\ \beta q\end{matrix};q,-xt\right){}_0\phi_1\left(\begin{matrix}-\\ \alpha q\end{matrix};q,-x\alpha qt\right).$$

Applying the Pfaff–Kummer transformation (2.2) to the ${}_2\phi_1$ appearing in the above generating function, we obtain

(3.1)

$$\sum_{n=0}^{\infty}\frac{q^{n(n-1)/2}t^n}{(\beta;q)_n(q;q)_n}\Phi_n^{\alpha,\beta}(x)=\frac{1}{(-xt;q)_\infty}{}_1\phi_1\left(\begin{matrix}\beta qx\\ \beta q\end{matrix};q,-t\right){}_0\phi_1\left(\begin{matrix}-\\ \alpha q\end{matrix};q,-x\alpha qt\right).$$

The singularities of the right-hand side of (3.1) are simple poles located at $t=-q^{-m}/x$, $m=0,1,\cdots$. Let

$$\Phi(x,t)=\sum_{n=0}^{\infty}\frac{q^{n(n-1)/2}t^n}{(\beta q;q)_n(q;q)_n}\Phi_n^{\alpha,\beta}(x). \tag{3.2}$$

Cauchy's theorem implies

$$\frac{q^{n(n-1)/2}}{(\beta q;q)_n(q;q)_n}\Phi_n^{\alpha,\beta}(x)=\frac{1}{2\pi i}\int_C \Phi(x,t)t^{-n-1}\,dt, \tag{3.3}$$

where C is a circle $|t|=\rho<1/|x|$. As we did in the case of Askey–Wilson polynomials, we think of C as a contour enclosing the point at infinity in the t-plane but with a clockwise orientation. Clearly

$$\begin{aligned}&\operatorname{Res}\{t^{-n-1}\Phi(x,t);t=-q^{-m}/x\}\\&\quad=-\frac{q^{mn}(-x)^n}{(q^{-m};q)_m(q;q)_\infty}{}_0\phi_1\left(\begin{matrix}-\\ \alpha q\end{matrix};q,\alpha q^{1-m}\right){}_1\phi_1\left(\begin{matrix}\beta qx\\ \beta q\end{matrix};q,q^{-m}/x\right).\end{aligned}$$

This and (3.3) establish the explicit formula

$$\begin{aligned}&\frac{(-x)^{-n}q^{n(n-1)/2}}{(\beta q;q)_n(q;q)_n}\Phi_n^{\alpha,\beta}(x)\\&\quad=\sum_{m=0}^{\infty}\frac{q^{mn+m(m+1)/2}}{(q;q)_m(q;q)_\infty}(-1)^m{}_0\phi_1\left(\begin{matrix}-\\ \alpha q\end{matrix};q,\alpha q^{1-m}\right){}_1\phi_1\left(\begin{matrix}-\\ \beta q\end{matrix};q,q^{-m}/x\right).\end{aligned} \tag{3.4}$$

This is also an asymptotic formula as $n\to\infty$ when x is fixed. Each of the hypergeometric functions appearing on the right-hand side of (3.4) is a sum of m terms. This can be seen as follows

$$\begin{aligned}{}_0\phi_1\left(\begin{matrix}-\\ \alpha q\end{matrix};q,\alpha q^{1-m}\right)&=\lim_{\varepsilon\to 0}{}_2\phi_1\left(\begin{matrix}\frac{1}{\varepsilon},\frac{1}{\varepsilon}\\ \alpha q\end{matrix};q,\alpha\varepsilon^2q^{1-m}\right)\\&=\lim_{\varepsilon\to 0}\frac{1}{(\alpha q;q)_\infty}{}_2\phi_1\left(\begin{matrix}q^{-m},\frac{1}{\varepsilon}\\ \alpha\varepsilon q^{1-m}\end{matrix};\varepsilon\alpha q\right)\end{aligned}$$

where we used the transformation, Askey and Ismail [4, p. 67],

$${}_2\phi_1\left(\begin{matrix}a,b\\ c\end{matrix};q,x\right)=\frac{(bx;q)_\infty(c/b;q)_\infty}{(x;q)_\infty(c;q)_\infty}{}_2\phi_1\left(\begin{matrix}\frac{abx}{c},b\\ bx\end{matrix};q,c/b\right). \tag{3.5}$$

This gives

$${}_0\phi_1\left(\begin{matrix}-\\ \alpha q\end{matrix};q,\alpha q^{1-m}\right)=\frac{1}{(\alpha q;q)_\infty}\sum_{k=0}^{m}\frac{(q^{-m};q)_k}{(q;q)_k}(-\alpha)^k q^{k(k+1)/2}.$$

Furthermore,

$$
{}_1\phi_1\left(\begin{matrix}\beta xq\\ \beta q\end{matrix};q,q^{-m}/x\right)=\lim_{\varepsilon\to 0}{}_2\phi_1\left(\begin{matrix}\frac{1}{\varepsilon},\beta xq\\ \beta q\end{matrix};q,\varepsilon q^{-m}/x\right)
$$

$$
=\frac{(1/x;q)_\infty}{(\beta q;q)_\infty}\,{}_2\phi_1\left(\begin{matrix}q^{-m},\beta xq\\ 0\end{matrix};q,1/x\right).
$$

Thus we proved

$$
(3.6)\quad \frac{(-x)^{-n}q^{n(n-1)/2}}{(\beta q;q)_n(q;q)_n}\Phi_n^{\alpha,\beta}(x)=\frac{(1/x;q)_\infty}{(\beta q;q)_\infty(\alpha q;q)_\infty}\sum_{m=0}^{\infty}\frac{q^{mn+m(m+1)/2}}{(q;q)_m(q;q)_\infty}(-1)^m
$$

$$
\cdot\,{}_2\phi_1\left(\begin{matrix}q^{-m},\beta xq\\ 0\end{matrix};q,1/x\right){}_1\phi_1\left(\begin{matrix}q^{-m}\\ 0\end{matrix};q,\alpha q\right).
$$

We now treat the big q-Jacobi polynomials. Ismail and Wilson [8] obtained the generating function

$$
\sum_{n=0}^{\infty}\frac{(\gamma q;q)_n t^n}{(q;q)_n(\beta q;q)_n}P_n(x)
$$

$$
=\left(\sum_{n=0}^{\infty}\frac{(\alpha q/x;q)_n(tx)^n}{(\alpha q;q)_n(q;q)_n}\right)\left(\sum_{k=0}^{\infty}\frac{(\beta x/\gamma;q)_k(-\gamma t)^k q^{k(k+1)/2}}{(q;q)_k(\beta q;q)_k}\right),
$$

that is,

$$
(3.7)\quad \sum_{n=0}^{\infty}\frac{(\gamma q;q)_n t^n}{(q;q)_n(\beta q;q)_n}P_n(x)={}_2\phi_1\left(\begin{matrix}0,\alpha q/x\\ \alpha q\end{matrix};q,xt\right){}_1\phi_1\left(\begin{matrix}\beta x/\gamma\\ \beta q\end{matrix};q,\gamma qt\right).
$$

Applying the transformation (2.2) to the ${}_2\phi_1$ in (3.7) gives

$$
(3.8)\quad \sum_{n=0}^{\infty}\frac{(\gamma q;q)_n t^n}{(\beta q;q)_n(q;q)_n}P_n(x)=\frac{1}{(xt;q)_\infty}{}_1\phi_1\left(\begin{matrix}x\\ \alpha q\end{matrix};\alpha qt\right){}_1\phi_1\left(\begin{matrix}\beta x/\gamma\\ \beta q\end{matrix};q\gamma qt\right).
$$

The poles of the right-hand side are all simple and are located at $t=q^{-m}/x$, $m=0,1,\cdots$. The residue of t^{-n-1} times the right-hand side of (3.8) at $t=q^{-m}/x$ is

$$
\frac{-x^n q^{mn}}{(q^{-m};q)_m(q;q)_\infty}{}_1\phi_1\left(\begin{matrix}x\\ \alpha q\end{matrix};q,\frac{\alpha}{x}q^{1-m}\right){}_1\phi_1\left(\begin{matrix}\beta x/\gamma\\ q\beta\end{matrix};\frac{\gamma}{x}q^{1-m}\right).
$$

This establishes

$$
(3.9)\quad \frac{(\gamma q;q)_n(q;q)_\infty}{(\beta q;q)_n(q;q)_n}P_n(x)=x^n\sum_{m=0}^{\infty}\frac{(-1)^m q^{mn+m(m+1)/2}}{(q;q)_m}
$$

$$
\cdot\,{}_1\phi_1\left(\begin{matrix}x\\ \alpha q\end{matrix};q,\frac{\alpha}{x}q^{1-m}\right){}_1\phi_1\left(\begin{matrix}\beta x/\gamma\\ q\beta\end{matrix};q,\frac{\gamma}{x}q^{1-m}\right).
$$

Each of the ${}_1\phi_1$'s appearing in (3.9) is a sum of m terms. This follows from (3.5) since

$$
(3.10)\quad {}_1\phi_1\left(\begin{matrix}\lambda\\ \mu\end{matrix};q,\frac{\mu}{\lambda}q^{-m}\right)=\lim_{\varepsilon\to 0}{}_2\phi_1\left(\begin{matrix}\frac{1}{\varepsilon},\lambda\\ \mu\end{matrix};\varepsilon\frac{\mu}{\lambda}q^{-m}\right)=\frac{(\mu/\lambda;q)_\infty}{(\mu q)_\infty}{}_2\phi_1\left(\begin{matrix}q^{-m},\lambda\\ 0\end{matrix};q,\frac{\mu}{\lambda}\right).
$$

This enables us to transform (3.9) to

(3.11)

$$\frac{(\theta;q)_n(q;q)_\infty}{(\beta q;q)_n(q;q)_n}P_n(x)=\frac{x^n(\gamma q;q)_\infty(\alpha q/x;q)_\infty}{(\alpha q;q)_\infty(\beta q;q)_\infty}\sum_{m=1}^{\infty}\frac{(-1)^m q^{mn+m(m+1)/2}}{(q;q)_m}\cdot\left(\begin{matrix}q^{-m},x\\0\end{matrix};q,\frac{\alpha q}{x}\right)_2\phi_1\left(\begin{matrix}q^{-m},\beta x/\gamma\\0\end{matrix};q,\frac{gq}{x}\right).$$

The relationship (3.11) is an explicit representation of $P_n(x)$. The right-hand side of (3.11) is clearly an asymptotic series as $n\to\infty$.

4. The continuous q-ultraspherical polynomials. The continuous q-ultraspherical polynomials $\{C_n(x;\beta\,|\,q)\}$ have the generating function

$$\sum_{n=0}^{\infty}C_n(\cos\theta,\beta\,|\,q)t^n=\frac{(\beta te^{i\theta};q)_\infty(\beta te^{-i\theta};q)_\infty}{(te^{i\theta};q)_\infty(te^{-i\theta};q)_\infty}. \tag{4.1}$$

They were discovered by L. J. Rogers around the turn of the century and Rogers used them to prove the Rogers–Ramanujan identities. Rogers did not realize that they are orthogonal polynomials although he was well aware of the fact that they generalize the ultraspherical (or Gegenbauer) polynomials. Their orthogonality was proved only recently, [4] and [6]. Recently, Rahman and Verma [10] proved the q-integral representation

$$C_n(\cos\theta,\beta\,|\,q)=\frac{2i\sin\theta(\beta;q)_\infty^2(\beta^2;q)_n}{(1-q)w_\beta(\cos\theta\,|\,q)(q;q)_\infty(\beta^2;q)_\infty}\cdot\int_{e^{i\theta}}^{e^{-i\theta}}u^n\frac{(que^{i\theta};q)_\infty(que^{-i\theta})_\infty}{(\beta ue^{i\theta};q)_\infty(\beta ue^{-i\theta};q)_\infty}d_qu, \tag{4.2}$$

where $0<\theta<\pi$, $w_\beta(x\,|\,q)$ is the weight function

$$w_\beta(\cos\theta\,|\,q)=(e^{2i\theta};q)_\infty(e^{-2i\theta};q)_\infty/[(\beta e^{2i\theta};q)_\infty(\beta e^{-2i\theta};q)_\infty],$$

and the q-integral is defined by

$$\int_0^a f(u)\,d_qu=a(1-q)\sum_{n=0}^{\infty}f(aq^n)q^n,\int_a^b f(u)\,d_qu=\int_0^b f(u)\,d_qu-\int_0^a f(u)\,d_qu.$$

The representation (4.2) resembles the familiar Dirichlet–Mehler formula, Szegö [12]. Rahman and Verma [10] observed that (4.2) can be used to compute the complete asymptotic expansion of $C_n(\cos\theta;\beta\,|\,q)$ for large n. They also used (4.2) to derive a new generating function for the continuous q-ultraspherical polynomials and to give a simple derivation of the Poisson kernel for the continuous q-ultraspherical polynomials that Gasper and Rahman obtained earlier [7].

Applying the procedure that we used in §§2 and 3 to the generating function (4.1), we obtain

$$C_n(\cos\theta;\beta\,|\,q)=\frac{(\beta;q)_\infty(\beta e^{2i\theta};q)_\infty}{(q;q)_\infty(e^{2i\theta};q)_\infty}e^{-in\theta}{}_2\phi_1\left(\begin{matrix}q/\beta,qe^{-2i\theta}/\beta\\qe^{-2i\theta}\end{matrix};q,\beta^2q^n\right)+\text{a similar term with }\theta\text{ replaced by }-\theta. \tag{4.3}$$

The relationship (4.3) is not exactly what Rahman and Verma obtained although it serves the same purpose. Applying the transformation

$$ {}_2\phi_1\left(\begin{matrix}a,b\\c\end{matrix};q,x\right)=\frac{(abx/c;q)_\infty}{(x;q)d\infty}\,{}_2\phi_1\left(\begin{matrix}c/a,c/b\\c\end{matrix};q,abx/c\right) $$

to the basic hypergeometric functions in (4.3) transforms (4.3) to

$$ (4.4)\quad C_n(\cos\theta;\beta\,|\,q)=\frac{(\beta;q)_\infty(\beta e^{2i\theta};q)_\infty(q^{n+1};q)_\infty}{(q;q)_\infty(e^{2i\theta};q)_\infty(\beta^2q^n;q)_\infty}\,{}_2\phi_1\left(\begin{matrix}\beta e^{2i\theta},\beta\\ qe^{-2i\theta}\end{matrix};q,q^{n+1}\right). $$

Formula (4.4) is a restatement of (4.2).

Acknowledgments. I thank Mizan Rahman and Arun Verma for sending me a preprint of their paper [10] which motivated this work. Discussions with Dennis Stanton of the University of Minnesota were extremely helpful.

REFERENCES

[1] G. E. ANDREWS, *On the q-analogue of Kummer's theorem and applications*, Duke Math. J., 40 (1973), pp. 525–528.

[2] G. E. ANDREWS AND R. A. ASKEY, *Classical orthogonal polynomials*, in Polynômes Orthogonaux et Applications, Proc., Bar-le-Duc, 1984, C. Brezinski et al, eds., Lecture Notes in Mathematics 1171, 1985, Springer Verlag, New York, pp. 36–62.

[3] _____, *q-analogues of the classical orthogonal polynomials and applications*, in preparation.

[4] R. A. ASKEY AND M. E. H. ISMAIL, *A generalization of ultraspherical polynomials*, in Studies in Pure Mathematics, P. Erdös, ed., Birkhauser, Basel, 1983, pp. 55–78.

[5] R. A. ASKEY AND J. A. WILSON, *A set of orthogonal polynomials that generalize the Racah coefficients or $6-j$ symbols*, this Journal, 10 (1979), pp. 1008–1016.

[6] _____, *Some basic hypergeometric orthogonal polynomials that generalize Jacobi polynomials*, Memoirs Amer. Math. Soc., 319 (1985).

[7] G. GASPER AND M. RAHMAN, *Positivity of the Poisson kernel for the continuous q-ultraspherical polynomials*, this Journal, 14 (1983), pp. 409–420.

[8] M. E. H. ISMAIL AND J. A. WILSON, *Asymptotic and generating relations for the q-Jacobi and ${}_4\phi_3$ polynomials*, J. Approx. Theory, 36 (1982), pp. 43–54.

[9] G. PONZANO AND T. REGGE, *Semi-classical limit of Racah coefficients*, in Spectroscopic and Group Theoretical Methods in Physics, North-Holland, Amsterdam, 1968, pp. 1–58.

[10] M. RAHMAN AND A. VERMA, *An integral representation of Rogers' q-ultraspherical polynomials*, Constructive Approx., to appear.

[11] L. J. SLATER, *Generalized Hypergeometric Functions*, Cambridge Univ. Press, Cambridge, 1966.

[12] G. SZEGÖ, *Orthogonal Polynomials*, fourth edition, Vol. 23, Colloquium Publications, American Mathematical Society, Providence, RI, 1975.

SIAM J. MATH. ANAL.
Vol. 17, No. 6, November 1986

016

MONOTONICITY PROPERTIES OF THE ZEROS OF BESSEL FUNCTIONS*

ÁRPÁD ELBERT† AND ANDREA LAFORGIA‡

Abstract. For $\nu \geqq 0$ let $c_{\nu k}$ be the kth positive zero of the general cylinder function

$$C_\nu(x) = \cos\alpha J_\nu(x) - \sin\alpha Y_\nu(x), \qquad 0 \leqq \alpha < \pi$$

where $J_\nu(x)$ and $Y_\nu(x)$ denote the Bessel functions of the first and second kind, respectively. Since the notation $c_{\nu k}$ does not reflect the dependence on the values of α, it is useful to define the function $j_{\nu\kappa}$ as in [3]. The sequence $j_{\nu 1} j_{\nu 2}, \ldots$ is used to denote the sequence of the zeros of $J_\nu(x)$ corresponding to $\alpha = 0$. Now for any $\kappa \in (k-1, k)$, where k is some natural number, let $j_{\nu\kappa} = c_{\nu k}$ with $\alpha = (k-\kappa)\pi$. The correspondence between $j_{\nu k}$ and $c_{\nu k}$ is one-to-one. In this paper we are concerned with some monotonicity properties related to $j_{\nu\kappa}$. We also study the convexity (concavity) of $j_{\nu\kappa}$ showing that $j_{\nu\kappa}$ is convex with respect to κ for fixed ν and $0 \leqq \nu \leqq \frac{1}{2}$, and concave for $\nu \geqq \frac{1}{2}$.

Finally for $\nu \geqq 0$ and $\kappa > 0$ we prove that the function $\log j_{\nu\kappa}$ is concave in ν for fixed κ and concave in κ for fixed ν.

Key words. zeros of Bessel functions

AMS(MOS) subject classification. Primary 33A40

1. Introduction.

For $\nu \geqq 0$ we use $c_{\nu k}$ to denote the kth positive zero of the cylinder function

$$C_\nu(x) = J_\nu(x)\cos\alpha - Y_\nu(x)\sin\alpha, \qquad 0 \leqq \alpha < \pi$$

where $J_\nu(x)$ and $Y_\nu(x)$ are the Bessel functions of the first and second kind, respectively.

The properties of $c_{\nu k}$ have been investigated by several authors.

Recently L. Lorch has studied the determinant [4, p. 223]

$$T = \begin{vmatrix} c_{\nu k} & c_{\nu+\delta, k+h} \\ c_{\nu+\varepsilon, k+r} & c_{\nu+\delta+\varepsilon, k+h+r} \end{vmatrix} \tag{1.1}$$

for $\varepsilon, \delta \geqq 0$, h, $r = 0, 1, 2, \cdots$, $\varepsilon + r > 0$, $h + \delta > 0$ and he has proved that $T < 0$.

In this paper we are interested, among other things, in determinants of the type (1.1), and we prove a more general result than the one given by Lorch.

In [3] we introduced the notation $j_{\nu\kappa}$ to denote the function $c_{\nu k}$ as follows: let $\kappa = k - \alpha/\pi$, then $j_{\nu\kappa} = c_{\nu k}$. Now we have $j_{\nu\kappa}$ for $\nu \geqq 0$, $\kappa > 0$ and with this notation (1.1) can be rewritten in the following way

$$T = \begin{vmatrix} j_{\nu\kappa} & j_{\nu+\delta, \kappa+h} \\ j_{\nu+\varepsilon, \kappa+r} & j_{\nu+\delta+\varepsilon, \kappa+h+r} \end{vmatrix} \tag{1.2}$$

where $\nu, \kappa, \delta, h, r, \varepsilon$ are nonnegative real numbers and $\varepsilon + r > 0$, $h + \delta > 0$. We shall see that Lorch's result is true also for T in (1.2).

* Received by the editors February 13, 1984, and in revised form August 8, 1985. This work was sponsored by the Consiglio Nazionale delle Ricerche, Italy.

† Mathematical Institute of the Hungarian Academy of Sciences, Budapest, P.f. 428, 1376 Hungary.

‡ Dipartimento di Matematica dell'Universitá di Torino, Via Carlo Alberto, 10, 10123 Torino, Italy.

Concerning the function $j_{\nu\kappa}$ we know that it is increasing with respect to both the variables. For fixed κ this follows from the Watson formula [6, p. 508]

$$\frac{d}{d\nu}j_{\nu\kappa}=2j_{\nu\kappa}\int_0^\infty K_0(2j_{\nu\kappa}\sinh t)e^{-2\nu t}dt \tag{1.3}$$

where $K_0(u)$ is the modified Bessel function of order zero, which is positive. On the other hand for fixed ν we have proved in [2] that $j_{\nu\kappa}$ is strictly increasing with respect to κ. For the sake of the later reference we express this property in the form

$$j_{\nu\kappa'}>j_{\nu\kappa}, \quad \kappa'>\kappa>0, \quad \nu>-\kappa. \tag{1.4}$$

Moreover we know [5] that the sequence $\{c_{\nu,k+1}-c_{\nu,k}\}_{k=1}^{\infty}$ is strictly decreasing for $|\nu|>\frac{1}{2}$ and increasing for $|\nu|<\frac{1}{2}$. So with our notation the sequence $\{j_{\nu,\kappa+k+1}-j_{\nu,\kappa+k}\}_{k=0}^{\infty}$ is monotonic. This property suggests that the function $j_{\nu\kappa}$ is concave with respect to κ if $\nu\geqq\frac{1}{2}$ and convex for $0\leqq\nu\leqq\frac{1}{2}$. Theorem 2.2 states that this observation is true.

2. The function $j_{\nu\kappa}$ and the determinant T. For $\nu\geqq 0$ and $\kappa>0$ we consider the derivative

$$\frac{\partial}{\partial\kappa}j_{\nu\kappa}=l_{\nu\kappa}=l. \tag{2.1}$$

Since in the special case $\nu=\frac{1}{2}j_{1/2,\kappa}=\kappa\pi$ [3], we have

$$l_{1/2,\kappa}=\pi. \tag{2.2}$$

Differentiating the function $dj_{\nu\kappa}/d\nu$ given by (1.3) with respect to κ, we obtain

$$\frac{d}{d\nu}l=2l\int_0^\infty K_0(2j\sinh t)e^{-2\nu t}dt$$
$$+2j\int_0^\infty K_0'(2j\sinh t)2l\sinh te^{-2\nu t}dt$$

where $j=j_{\nu\kappa}$. An integration by parts of the second integral on the right-hand side gives

$$2l\int_0^\infty K_0'(2j\sinh t)2j\cosh t\tanh te^{-2\nu t}dt=2l\left[K_0(2j\sinh t)\tanh te^{-2\nu t}\right]_0^\infty$$
$$-2l\int_0^\infty K_0(2j\sinh t)\left[\frac{1}{\cosh^2 t}-2\nu\tanh t\right]e^{-2\nu t}dt.$$

Recalling that

$$K_0(u)=\begin{cases}O\left(\log\dfrac{1}{u}\right), & u>0, \quad u\to 0,\\ o(e^{-u}), & u\to\infty,\end{cases}$$

the first term on the right-hand side is zero; therefore by (2.1)

$$\frac{d}{d\nu}l=2l\int_0^\infty K_0(2j\sinh t)(\tanh^2 t+2\nu\tanh t)e^{-2\nu t}dt.$$

Taking into account (2.2), we have

$$(2.3)\qquad l_{\nu\kappa}=\pi\exp\left\{2\int_{1/2}^{\nu}\left[\int_0^{\infty}K_0(2j_{\mu\kappa}\sinh t)(\tanh^2 t+2\mu\tanh t)e^{-2\mu t}dt\right]d\mu\right\}.$$

A simple consequence of the formula (2.3) is the following result.

THEOREM 2.1. *For $\nu\geqq 0$ and $\kappa'>\kappa>0$ the relations*

$$j_{\nu\kappa'}-j_{\nu\kappa}\begin{cases}>\pi(\kappa'-\kappa), & \nu>\frac{1}{2},\\ =\pi(\kappa'-\kappa), & \nu=\frac{1}{2},\\ <\pi(\kappa'-\kappa), & 0\leqq\nu<\frac{1}{2}\end{cases}$$

hold.

Proof. By (2.3) we have the inequalities $l_{\nu\kappa}\gtreqless\pi$ if $\nu\gtreqless\frac{1}{2}$ respectively, and $j_{\nu\kappa'}-j_{\nu\kappa}=\int_{\kappa}^{\kappa'}l_{\nu\lambda}d\lambda$; hence the conclusion of the Theorem follows.

THEOREM 2.2. *The function $j_{\nu\kappa}$ is concave with respect to κ if $\nu\geqq\frac{1}{2}$ and convex if $0\leqq\nu\leqq\frac{1}{2}$.*

Proof. To show the concavity of the function $j_{\nu\kappa}$ with respect to κ, in the case $\nu\geqq\frac{1}{2}$, we need to show that the function $l_{\nu\kappa}=\partial j_{\nu\kappa}/\partial\kappa$ decreases as κ increases. Let $\kappa'>\kappa>0$. Then by (1.4) $j_{\nu\kappa'}>j_{\nu\kappa}$ for $\nu\geqq 0$. On the other hand the function $K_0(u)$ has the integral representation [6, p. 446]

$$K_0(u)=\int_0^{\infty}e^{-u\cosh\xi}d\xi;$$

hence $K_0(u)$ strictly decreases as u increases and by (2.3) we have $l_{\nu\kappa}>l_{\nu\kappa'}$. This gives the concavity of $j_{\nu\kappa}$ with respect to κ, for $\nu\geqq\frac{1}{2}$.

Similarly we can prove the convexity of $j_{\nu\kappa}$ with respect to κ, in the case $0\leqq\nu\leqq\frac{1}{2}$, and the proof of Theorem 2.2 is complete.

THEOREM 2.3. *For $\nu\geqq 0$ and $\kappa>0$ the function $\log j_{\nu\kappa}$ is concave in ν for fixed κ and concave in κ for fixed ν.*

Proof. The first part of the theorem follows from

$$\frac{j_{\nu+\varepsilon+\delta,\kappa}}{j_{\nu+\varepsilon,\kappa}}<\frac{j_{\nu+\delta,\kappa}}{j_{\nu\kappa}},\qquad \varepsilon,\delta>0;$$

thus we need to show that the function $H(\nu,\kappa)$ defined by

$$(2.4)\qquad H(\nu,\kappa)=\frac{j_{\nu+\delta,\kappa}}{j_{\nu\kappa}}$$

decreases as ν increases. By (1.3) we have

$$(2.5)\qquad \log H(\nu,\kappa)=2\int_{\nu}^{\nu+\delta}\left[\int_0^{\infty}K_0(2j_{\mu\kappa}\sinh t)e^{-2\mu t}dt\right]d\mu$$

and similarly

$$\begin{aligned}\log H(\nu+\varepsilon,\kappa)&=2\int_{\nu+\varepsilon}^{\nu+\varepsilon+\delta}\left[\int_0^{\infty}K_0(2j_{\mu\kappa}\sinh t)e^{-2\mu t}dt\right]d\mu\\&=2\int_{\nu}^{\nu+\delta}\left[\int_0^{\infty}(2j_{\mu+\varepsilon,\kappa}\sinh t)e^{-2(\mu+\varepsilon)t}dt\right]d\mu.\end{aligned}$$

By (1.3) $j_{\mu+\varepsilon,\kappa}>j_{\mu,\kappa}$ and since $K_0(u)$ decreases as u increases, the function in the last integral is less than the one in (2.5). This shows that $H(\nu+\varepsilon,\kappa)<H(\nu,\kappa)$ and the result follows.

For the proof of the second part we have to show the inequality

$$\frac{j_{\nu,\kappa+r+h}}{j_{\nu,\kappa+r}}<\frac{j_{\nu,\kappa+h}}{j_{\nu\kappa}}, \qquad r,h>0$$

which is equivalent to proving that the function $L(\nu,\kappa)$ defined by

$$L(\nu,\kappa)=\frac{j_{\nu,\kappa+h}}{j_{\nu\kappa}} \tag{2.6}$$

decreases as κ increases.

By (2.1) and (2.6) we have

$$\frac{\partial}{\partial\kappa}\log L(\nu,\kappa)=\frac{l_{\nu,\kappa+h}}{j_{\nu,\kappa+h}}-\frac{l_{\nu\kappa}}{j_{\nu\kappa}} \tag{2.7}$$

and we need to prove that $l_{\nu\kappa}/j_{\nu\kappa}$ decreases as κ increases. In the particular case $\nu=\frac{1}{2}$ this is trivial, because by (2.2) $l_{1/2,\kappa}=\pi$ and $j_{1/2,\kappa}=\kappa\pi$.

Now we distinguish the cases $\nu>\frac{1}{2}$ and $0\leqq\nu\leqq\frac{1}{2}$. In the first case for $r>0$ and $\kappa'=\kappa+r$ we have by (1.4) and (2.3)

$$\frac{l_{\nu\kappa'}}{j_{\nu\kappa'}}<\frac{l_{\nu\kappa'}}{j_{\nu\kappa}}<\frac{l_{\nu\kappa}}{j_{\nu\kappa}},$$

so by (2.7) the function $L(\nu,\kappa)$ decreases as κ increases, provided $\nu>\frac{1}{2}$.

The case $0\leqq\nu\leqq\frac{1}{2}$ requires more detailed investigation. By (2.1) and (2.3) we get

$$\psi(\nu,\kappa)=\frac{\partial}{\partial\kappa}\log\frac{l_{\nu\kappa}}{j_{\nu\kappa}}=-2\int_\nu^{1/2} \cdot\left[\int_0^\infty K_0'(2j_{\mu\kappa}\sinh t)2l_{\mu\kappa}\sinh t(\tanh^2 t+2\mu\tanh t)e^{-2\mu t}dt\right]d\mu-\frac{l_{\nu\kappa}}{j_{\nu\kappa}}. \tag{2.8}$$

We claim that $\psi(\nu,\kappa)<0$ for $0\leqq\nu\leqq\frac{1}{2}$. For $\nu=\frac{1}{2}$ this is trivial. We show that $\partial\psi(\nu,\kappa)/\partial\nu>0$ for $\nu>0$. Indeed by (1.3), (2.3) and (2.8) we get

$$\frac{j_{\nu\kappa}}{l_{\nu\kappa}}\frac{\partial}{\partial\nu}\psi(\nu,\kappa)=I_1+2\nu I_2 \tag{2.9}$$

where

$$I_1=4j_{\nu\kappa}\int_0^\infty K_0'(2j_{\nu\kappa}\sinh t)\sinh t\tanh^2 t\,e^{-2\nu t}dt$$

$$+2\int_0^\infty K_0(2j_{\nu\kappa}\sinh t)(1-\tanh^2 t)e^{-2\nu t}dt,$$

and

$$I_2=\int_0^\infty\left[4j_{\nu\kappa}\sinh t\tanh tK_0'(2j_{\nu\kappa}\sinh t)-2\tanh tK_0(2j_{\nu\kappa}\sinh t)\right]e^{-2\nu t}dt.$$

An integration by parts gives for $\nu > 0$

$$I_2 = \left\{ \left[4 j_{\nu\kappa} \sinh t \tanh t K_0'(2 j_{\nu\kappa} \sinh t) - 2 \tanh t K_0(2 j_{\nu\kappa} \sinh t) \right] \frac{e^{-2\nu t}}{-2\nu} \right\}_0^\infty$$
$$+ \frac{1}{2\nu} \int_0^\infty \left[4 j_{\nu\kappa} \left(\sinh t + \frac{\sinh t}{\cosh^2 t} \right) K_0'(2 j_{\nu\kappa} \sinh t) \right.$$
$$+ 8 j_{\nu\kappa}^2 \sinh^2 t K_0''(2 j_{\nu\kappa} \sinh t) - \frac{2}{\cosh^2 t} K_0(2 j_{\nu\kappa} \sinh t)$$
$$\left. - 4 j_{\nu\kappa} \sinh t K_0'(2 j_{\nu\kappa} \sinh t) \right] e^{-2\nu t} dt.$$

Recalling that $K_0'(u) = -K_1(u)$ and

$$K_1(u) = \begin{cases} O\left(\log \frac{1}{u}\right), & u > 0, \quad u \to 0, \\ o(e^{-u}), & u \to \infty \end{cases}$$

the first term on the right-hand side is zero and by (2.3) we obtain

(2.10)

$$\frac{j_{\nu\kappa}}{l_{\nu\kappa}} \frac{\partial}{\partial \nu} \psi(\nu, \kappa) = 2 \int_0^\infty \left[T_1(t) K_0(2 j_{\nu\kappa} \sinh t) + 2 j_{\nu\kappa} T_2(t) K_0'(2 j_{\nu\kappa} \sinh t) \right.$$
$$\left. + 4 j_{\nu\kappa}^2 \sinh^2 t K_0''(2 j_{\nu\kappa} \sinh t) \right] e^{-2\nu t} dt$$

where

$$T_1(t) = 1 - \tanh^2 t - \frac{1}{\cosh^2 t} = 0,$$

$$T_2(t) = \sinh t \tanh^2 t + \frac{\sinh t}{\cosh^2 t} = \sinh t.$$

Taking into account that $K_0(u)$ satisfies the differential equation [1, p. 374]

$$u^2 K'' + u K' - u^2 K = 0,$$

we have

$$\frac{\partial}{\partial \nu} \psi(\nu, \kappa) = 8 l_{\nu\kappa} j_{\nu\kappa} \int_0^\infty K_0(2 j_{\nu\kappa} \sinh t) \sinh^2 t e^{-2\nu t} dt, \tag{2.11}$$

which is clearly positive. So the function $\psi(\nu, \kappa)$ increases as ν increases and by (2.8)

$$\psi(\nu, \kappa) = \frac{\partial}{\partial \kappa} \log \frac{l_{\nu\kappa}}{j_{\nu\kappa}} \leqq \psi\left(\frac{1}{2}, \kappa\right) = \frac{-1}{\kappa} < 0, \qquad 0 \leqq \nu \leqq \tfrac{1}{2}.$$

Consequently the function $l_{\nu\kappa}/j_{\nu\kappa}$ decreases as κ increases and by (2.7) $L(\nu, \kappa)$ decreases as κ increases. The proof of Theorem 2.3 is complete.

COROLLARY 2.1. *For $\varepsilon, \delta, h, r > 0$ let T be defined by* (1.2). *If $\varepsilon + r > 0$ and $h + \delta > 0$, then $T < 0$.*

Proof. The inequality $T<0$ is equivalent to

$$H(\nu+\varepsilon,\kappa+r)L(\nu+\delta+\varepsilon,\kappa+r)<H(\nu,\kappa)L(\nu+\delta,\kappa), \tag{2.12}$$

where the functions $H(\nu,\kappa)$ and $L(\nu,\kappa)$ have been defined by (2.4) and (2.6), respectively.

So it is sufficient to show that the functions $H(\nu,\kappa)$ and $L(\nu,\kappa)$ are decreasing with respect to both the variables.

Concerning the function $H(\nu,\kappa)$ we have already shown that it is decreasing with respect to ν.

Now let us suppose $r>0$ and $\kappa'=\kappa+r$. Then by (1.4) and (2.5)

$$\log H(\nu,\kappa')=2\int_{\nu}^{\nu+\delta}\left[\int_0^{\infty}K_0(2j_{\mu\kappa'}\sinh t)e^{-2\mu t}dt\right]d\mu$$

$$<2\int_{\nu}^{\nu+\delta}\left[\int_0^{\infty}K_0(2j_{\mu\kappa}\sinh t)e^{-2\mu t}dt\right]d\mu=\log H(\nu,\kappa).$$

This shows that the function $H(\nu,\kappa)$ decreases as κ increases, provided $\delta>0$.

Let us consider the function $L(\nu,\kappa)$ defined by (2.6). The decrease of L as a function of κ has been shown. So we have to prove that $L(\nu+\delta,\kappa)<L(\nu,\kappa)$ for $\delta>0$, or equivalently

$$\frac{j_{\nu+\delta,\kappa+h}}{j_{\nu,\kappa+h}}<\frac{j_{\nu+\delta,\kappa}}{j_{\nu,\kappa}}, \qquad \delta>0.$$

By (2.4) this inequality is equivalent to $H(\nu,\kappa+h)<H(\nu,\kappa)$, which is the property proved above.

We have seen that the inequality (2.12) is equivalent to $T<0$; hence we need the sharp inequality in (2.12). Since $\varepsilon+r>0$ and $\varepsilon\geqq 0$, $r\geqq 0$, there are two possibilities: the first one, $r>0$, $\varepsilon\geqq 0$ and the second one $r=0$, $\varepsilon>0$. In the first case we have $L(\nu+\delta+\varepsilon,\kappa+r)<L(\nu+\delta,\kappa+r)<L(\nu+\delta,\kappa)$ and $H(\nu+\varepsilon,\kappa+r)\leqq H(\nu,\kappa)$ and (2.12) holds. In the second case we have the sharp inequality $H(\nu+\varepsilon,\kappa)<H(\nu,\kappa)$, and hence $T<0$. The proof of the Corollary 2.1 is complete.

Acknowledgment. We are grateful to the referee, whose constructive criticism and suggestions led to the present improved version of this paper.

REFERENCES

[1] M. ABRAMOWITZ AND I. A. STEGUN, eds., *Handbook of Mathematical Functions*, Dover, New York, 1970.

[2] Á. ELBERT AND A. LAFORGIA, *On the convexity of zeros of Bessel functions*, this Journal, 10 (1985), pp. 614–619.

[3] _____, *On the square of zeros of Bessel functions*, this Journal, 15 (1985), pp. 206–212.

[4] L. LORCH, *Turánians and Wronskians for the zeros of Bessel functions*, this Journal, 11 (1980), pp. 223–227.

[5] C. H. STURM, *Mémoire sur les équations differentielles du second ordre*, J. Math. Pures Appl., 1 (1836), pp. 106–186.

[6] G. N. WATSON, *A Treatise on the Theory of Bessel Functions*, 2nd ed., Cambridge Univ. Press, Cambridge, 1944.

SIAM J. MATH. ANAL.
Vol. 17, No. 6, November 1986

017

THE SUMMABILITY SIGNIFICANCE OF ROOTS OF BESSEL FUNCTIONS*

E. C. OBI†

Abstract. A new semicontinuous Toeplitz (regular) matrix method of summability is developed from the (roots of) Bessel functions. Like the method of the Bessel function itself, developed by R. G. Cooke, viz.,

$$(J,v): a_{\lambda k} = 2J_{k+\nu}^2(\lambda),$$

the new method,

$$(\mathcal{O},m): t_{\nu k}^{(m)} = \beta(m,\nu)\, j_{\nu k}^{-2m}$$

(where $\beta(m,\nu)$ is explicit in m, ν), falls into the "Cesàro scope," but unlike the Cooke method, the $(\mathcal{O},m)$ limitation methods are definitely consistent with $(C,1)$ throughout the convergence field of this Cesàro means.

1. Introduction. The semicontinuous matrix of R. G. Cooke,

$$a_{\lambda k} = 2J_{k+\nu}^2(\lambda) \qquad (\lambda > 0,\ k = 1,2,\cdots),$$

where ν is fixed in $R \equiv (-\infty, \infty)$, is a Toeplitz (i.e. regular) method of summability (cf. [1]). A sequence, $z \equiv \{z_k\}$ in C (the set of complex numbers), is said to be (J,ν)-summable to $\gamma \in C$ if

$$\lim_{\lambda \to \infty} \sum_{k=1}^{\infty} 2J_{k+\nu}^2(\lambda) z_k = \gamma.$$

Thus, for example, summability $(J,0)$ would be consistent with $(C,1)$-summability for such divergent sequences as $z_k = \cos k\theta$ if $\theta \notin \{2n\pi : n = 0, \pm 1, \pm 2, \cdots\}$ is fixed (cf. [1]). The $(J, \frac{1}{2})$-method sums divergent Fourier series to their functions, a.e. on $[-\pi, \pi]$. With these and other related results of Cooke the summability significance of $J_\nu(z)$ was brought to light. Some cylinder (or their related) functions have played other classical roles in the theory of infinite matrices. It is of interest to note (in §3) that the zeros, $j_{\nu k}$, of the particular cylinder functions, $J_\nu(z)$, have a strong relationship with an important class of summability methods, viz. the *delicate* class. (We recall that there is the powerful class (cf. [2, p. 197]), consisting of those methods which are efficient for crudely divergent series—such as a power series outside the disc of convergence—but are usually ineffective for the slowly oscillating types, such as $\sum_{k=1}^{\infty}(e^{i\sqrt{k}}/\sqrt{k})$ [2, p. 197].) A summability method which is efficient for the slowly oscillating types, but normally not for the former, is said to belong to the delicate class.

If we identify the element $(\nu,k) \in R^+ \times N \equiv [0,\infty) \times \{1, 2, \cdots\}$ by the kth zero of $J_\nu(z)$, we construct explicit functions $\beta(m,\nu)$ such that for each fixed $m \in N$, the transformation, $t^{(m)}: R^+ \times N \to R^+$, defined by

$$t^{(m)}: j_{\nu k} \to \beta(m,\nu)\, j_{\nu k}^{-2m},$$

transforms the nonregular matrix $j_{\nu k}$ to a regular semicontinuous method of summability which belongs to the delicate class and which, in addition, is endowed with other desirable consistency properties. In particular, this matrix-construct of $j_{\nu k}$ is efficient

*Received by the editors May 3, 1983, and in revised form November 1, 1984.
† Department of Mathematics, University of Nigeria, Nsukka, Nigeria.

for Fourier series and has a more favorable comparison with $(C,1)$ than was shown (Cooke [1]) between (J,ν) and $(C,1)$.

2. Summability theory and roots of $J_\nu(z)$. Denote $\mathscr{S}\equiv$ the linear space of all complex sequences, $\mathbb{C}\equiv\{z\in\mathscr{S}:\lim z \text{ exists}\}$, and $\mathscr{H}\equiv$ the set of all divergent sequences of 0's and 1's. If $M\equiv(m_{\nu k})$, $\nu\geqq\nu_0>0$, $k=1,2,\cdots$, is an infinite semicontinuous complex matrix such that

$$\omega(\nu)\equiv\sum_{k=1}^{\infty}m_{\nu k}z_k \tag{1}$$

exists for all $\nu\geqq\nu_0$, then the M-transforms $\{\omega(\nu)\}_{\nu\geqq\nu_0}$ of $z\equiv\{z_k\}$ may be denoted by $w=Mz$ or by (1). Let $\mathscr{C}(M)$ be the *convergence field* of M (i.e. the linear space of those $z\in\mathscr{S}\ni\lim Mz$ exists), and let $\mathscr{P}(M)$ be the subspace of $\mathbb{C}\cap\mathscr{C}(M)$ consisting of those $z\ni\lim z=\lim Mz$. If $\mathbb{C}\subseteq\mathscr{C}(M)$, M is called a *K-matrix* (after Kojima [2, p. 4]). Those K-matrices that satisfy the following *consistency* or *regularity* property, $\mathscr{P}(M)=\mathbb{C}$, are called *Toeplitz* (or consistent, or regular, or simply T-) methods of summability, or just T-matrices. The *M-generalized limit* of z is $\lim Mz\equiv\gamma$ if γ exists and if M is Toeplitz. For a proof that the semicontinuous matrix. $a_{xk}=2J^2_{k+\nu}(x)$, with ν fixed in R, and $k\in N$, and x varying in $(0,\infty)$, is Toeplitz (regular), see [1].

Two among the best known methods which we shall have an occasion to use below are the Cesàro and the Riesz means:

$$C_{(r)}:c_{ik}=\begin{cases}\dfrac{A^{r-1}_{i-k}}{A^r_{i-1}}, & \text{if } 1\leqq k\leqq i,\\ 0 & \text{if } k>i,\ i,k=1,2,\cdots, \text{ where } A^r_0=1,\end{cases}$$

$A^r_i=(r+1)\cdots(r+i)/i!$, if $i\geqq1$ (r is real but not a negative integer); the corresponding transformation is

$$W(i)=\left(A^r_{i-1}\right)^{-1}\sum_{k=1}^{i}A^{r-1}_{i-k}z_k$$

(the notation $C_{(r)}$ and (C,r) will be used interchangeably according to convenience);

$$(\mathscr{R},r):r_{ik}=\begin{cases}\left(1-\dfrac{k}{i}\right)^r-\left(1-\dfrac{k+1}{i}\right)^r & \text{if } i-k>1,\\ 0 & \text{otherwise.}\end{cases}$$

The M-sum of a series will mean that of the sequence of its partial sums, the quantities $\lim_{\nu\to\infty}m_{\nu k}$ and $\lim_{\nu\to\infty}\Sigma_k m_{\nu k}$ are called the (first and second) characteristic numbers of M; and finally $[x]$ will denote the integral part of x. Now, using the zeros of Bessel functions, we construct a new J_ν-related regular method of summability which is of the delicate class and has the properties suggested in §1. It will be useful to define the following quantities:

$$e_m=\text{the } m\text{th Catalan number } \frac{1}{m}\binom{2m-2}{m-1};$$

$$d(m)=1-2m+\sum_{s=1}^{m}[m/s];\ h_m(\nu)=e_m(\nu+m-1)^{d(m)};$$

$$\pi_m(\nu)=4^m\sum_{s=1}^{m}(\nu+s)^{[m/s]},$$

for all $m\geqq1$, $\nu\geqq0$.

THEOREM 1. *If $j_{\nu k}$ is the kth positive root of $J_\nu(z)$, $\nu \in R^+$, $k=1,2,3,\cdots$, then each of the semicontinuous matrices generated below by $j_{\nu k}$ according to the equations,*

$$t_{\nu k}^{(m)} = \pi_m(\nu)\{h_m(\nu)\}^{-1} j_{\nu k}^{-2m}, \qquad m=1,2,3,\cdots,$$

is a Toeplitz (regular) method of summability.

(The symbol $(\mathcal{O},m)$, or $\mathcal{O}$ if m is fixed, will below denote this method, and m will be called its degree or order.)

Before the proof we first discuss a result which will be needed from the theory of Bessel functions.

Consider the Maclaurin coefficients, $\sigma_m(\nu)=\sum_{k=1}^{\infty} j_{\nu k}^{-2m}$, of the function, $\sqrt{z}\,J_{\nu+1}(\sqrt{z})/(2J_\nu(\sqrt{z}))=\sum_{m=1}^{\infty}\sigma_m(\nu)z^m$ (cf. [3, p. 528]), where for our purposes $\nu \geqq 0$. Then the product,

$$\phi_m(\nu) \equiv \left\{4^m \prod_{s=1}^{m} (\nu+s)^{[m/s]}\right\}\sigma_m(\nu),$$

is by a theorem of Kishore [3a][1] a polynomial in ν which can be represented in the form,

$$\phi_m(\nu) = \sum_{i=1}^{c(m)} 2^{m_i} \prod_{s=2}^{m-1} (\nu+s)^{a_{is}} (m \geqq 3), \qquad \phi_1=\phi_2=1, \tag{2}$$

where

$$c(m) = \sum_{s=1}^{[m/2]} c(s)c(m-s), \qquad c(1)=1,$$

$$\sum_{i=1}^{c(m)} 2^{m_i} = \frac{1}{m}\binom{2m-2}{m-1}, \quad \text{and} \quad \sum_{s=2}^{m-1} a_{is} = \deg(\phi_m) = 1-2m+\sum_{s=1}^{m}[m/s]. \tag{2a}$$

The proof of (2) may be found in [3a, p. 515].[1]

Thus $\sigma_m(\nu)$ is rational. Historically it was first used by Lord Rayleigh in his search for the numerical value for $\min_k j_{\nu k}$ when $\nu \geqq 0$ [4, p. 502].

Proof of Theorem 1. We verify that $(\mathcal{O},m)$ satisfies the Silverman–Toeplitz conditions for regularity by showing that there exists $\lambda>0$, independent of ν in $0 \leqq \nu < \infty$, such that $\sum_k t_{\nu k}^{(m)} \leqq \lambda$, and that the (first and second) characteristic numbers of $(\mathcal{O},m)$ are 0 and 1, respectively. Now take the polynomial ϕ_m. Since $\sum_{s=2}^{m-1} a_{is} = d(m)$ from (2a), we can write (2) as

$$\phi_m(\nu) = \sum_{i=1}^{c(m)} 2^{m_i} \prod_{s=1}^{d(m)} (\nu+b_{is}),$$

for some integers b_{is}, $2 \leqq b_{is} \leqq m-1$. Then on $\nu \geqq 0$,

$$0 \leqq \phi_m(\nu) \leqq (\nu+m-1)^{d(m)} \sum_{i=1}^{c(m)} 2^{m_i} = (\nu+m-1)^{d(m)} e_m = h_m(\nu);$$

[1] The notation ϕ_{2n} reads as ϕ_n after 1964.

and so,

$$\sum_k t_{\nu k}^{(m)} = \pi_m(\nu)\{h_m(\nu)\}^{-1}\sum_k j_{\nu k}^{-2m} = \pi_m(\nu)\{h_m(\nu)\}^{-1}\sigma_m(\nu)$$

$$= \phi_m(\nu)\{h_m(\nu)\}^{-1} \leqq 1 \quad \forall \nu \geqq 0.$$

Thus the first Silverman–Toeplitz condition holds for $(\mathcal{O}, m)$D with $\lambda = 1$. Next, since $j_{0k} > 2$ for all $k \geqq 1$, then using the MacCann inequality [5, pp. 101–2],

$$j_{\nu k}^2 \geqq j_{0k}^2 + \nu^2, \quad j_{\nu k}^2 > \pi^2\left(k - \tfrac{1}{4}\right)^2 + \nu^2 \quad (\nu \geqq 0), \tag{3}$$

we find that

$$0 < t_{\nu k}^{(m)} = \pi_m(\nu)\{h_m(\nu)\}^{-1} j_{\nu k}^{-2m} < 4^m \prod_{s=1}^{m} (\nu + s)^{[m/s]} / \{h_m(\nu)(4+\nu^2)^m\} \equiv f(\nu)/g(\nu),$$

where f, g are polynomials in ν, with $\deg(f) = \sum_{s=1}^m [m/s]$, and $\deg(g) = d(m) + 2m = 1 + \sum_{s=1}^m [m/s] > \deg(f)$. Therefore, $\lim_{\nu \to \infty} f(\nu)/g(\nu) = 0$, whence $\lim_{\nu \to \infty} t_{\nu k}^{(m)} = 0$.

Finally, write $\sum_k t_{\nu k}^{(m)} = \phi_m(\nu)\{h_m(\nu)\}^{-1}$, which we obtained above. Since ϕ_m and h_m have the same degrees in ν, and the same leading coefficients, e_m, it follows that $\sum_k t_{\nu k}^{(m)} \to 1$, as $\nu \to \infty$. This completes the proof. □

Now, regular methods A, P will be said to be *comparable* at $z = \{z_k\} \in \mathcal{S}$, if $z \in \mathcal{C}(A) \cap \mathcal{C}(P)$. If A, P are comparable at z, and A-sum $= P$-sum at z, we say that A and P are *consistent at* z. Thus consistency on a set X will mean consistency at all $z \in X$. On the other hand, we will describe a regular method A as being *consistent with* P at z if P-summability of $z \Rightarrow A$-summability of z to the same sum. (In case A is consistent with P everywhere in $\mathcal{S}$, we just say, "A is consistent with P." Example: if $r, r' > 0$, (C, r) and (C, r') are consistent on $\mathcal{C}(C, r) \cap \mathcal{C}(C, r')$; $(C, 2)$ is consistent with $(C, 1)$ but not vice versa. Otherwise we will specify the subset X of $\mathcal{S}$ to which the phrase refers.) Next we write, "$A \sim P$ at z" (i.e. equivalent at z), to mean that *either* both A-sum and P-sum of z exist and are equal, *or else* neither exists. If (a) $A \sim P$ at z, and (b) $Az - Pz$ is a null sequence (in case neither A-sum nor P-sum exists at z), we say that A is *absolutely equivalent* to P at z, and we write, "$A \equiv P$ at z." The example, $(C, r) \equiv (\mathcal{R}, r)$ on all of l^∞, when $r > 0$, is well known.

Once a K-method, A, is proved regular, it raises other natural and important classical questions. Does it have the Borel property?[2] Where is it consistent with some older method? What divergent sequences may it sum to their right values? What about the summability of such important series as divergent Fourier series? If $l^\infty(A)$ is the proper subspace of the nonseparable metric space l^∞ (with the usual sup metric), consisting of all the A-summable elements of l^∞, is $l^\infty(A)$ also nonseparable? And so on. Answers to these questions as they affect $(\mathcal{O}, m)$ are taken up next.

3. An efficiency theorem. Let two regular methods A, P be related as follows: (a) A is consistent with P; and (b) the sequence Az is bounded whenever Pz is bounded. Then A is said to be at least as *efficient* (or at least as *powerful*) as P. We claim that every $(\mathcal{O}, m)$ is at least as efficient as $(C, 1)$ summability.

LEMMA 1. *If $P = (p_{ik})$ is a T matrix such that for each fixed i, $kp_{ik} \to 0$, as $k \to \infty$, then*

(a) $\sum_k p_{ik} = \sum_k b_{ik}$, *where* $b_{ik} = k(p_{ik} - p_{i,k+1})$, *and*

(b) $\sum_k p_{ik} z_k = \sum_k b_{ik} S_k / k$,

[2] Does the A-transform of *almost* all divergent sequences of 0's and 1's converge to $\frac{1}{2}$? This important property is held by many well-known regular methods, but is not necessary for regularity. For a definition of "almost," see for example [2, p. 207].

whenever $z=\{z_k\}\in\mathscr{S}$ *is such that*

$$\frac{1}{k}S_k\equiv\frac{1}{k}\sum_{n=1}^{k}z_n$$

is a bounded sequence. (For this known fact, cf. [2, p. 92, no. 5].)

THEOREM 2. *The* $\mathcal{O}$*-method,* $t_{ik}^{(m)}$*, of any degree* $m\geqq 1$*, is at least as efficient as the Cesàro means* $(C,1)$ *as a method of summability.*

Proof. Fix $m\geqq 1$, and denote $\mathcal{O}\equiv(\mathcal{O},m)$ and $t_{ik}\equiv t_{ik}^{(m)}$. Now for every i fixed,

$$0<kt_{ik}=k\Pi_m(i)\{h_m(i)\}^{-1}j_{ik}^{-2m}<\Pi_m(i)\{h_m(i)\}^{-1}\pi^{-2m}k\left(k-\frac{1}{4}\right)^{-2m}\to 0$$

as $k\to\infty$, in view of the MacCann inequality (3). So,

$$\lim_k kt_{ik}=0. \tag{4}$$

Next, set $b_{ik}=k(t_{ik}-t_{i,k+1})$. Then with i fixed, $t_k\equiv t_{ik}$ is a strictly decreasing sequence of positive numbers, so that $b_{ik}>0$, $\forall k$. Since $t_{ik}>0$ is a T-matrix, $\exists\lambda>0$, independent of i, such that in view of (4) and Lemma 1(a),

$$0<\sum_k b_{ik}=\sum_k t_{ik}\leqq\lambda.$$

It is now clear that b_{ik} is a T-matrix also. Now let $\{\omega(i)\}_{i\geqq i_0}$ be the $\mathcal{O}$-transform, $\mathcal{O}_z$, of $z=\{z_k\}$, and suppose $C_{(1)}z$ is bounded. Then $(1/k)S_k$ is a bounded sequence (say by α). Therefore,

$$\omega(i)=\sum_k t_{ik}z_k=\sum_k b_{ik}S_k/k \tag{5}$$

(by Lemma 1(b)), and so $|\omega(i)|\leqq\alpha\lambda$, $\forall i$; hence $\mathcal{O}z$ is bounded.

For the consistency aspect, suppose $C_{(1)}z\to\gamma\in C$, for a given $z\in\mathscr{S}$. Then ($(1/k)S_k\to\gamma$ as $k\to\infty$, so that) (5) applies. Hence, since b_{ik} is a T-matrix, and $(1/k)S_k\to\gamma$, we must have $\omega(i)\to\gamma$ as $i\to\infty$. That is, $\lim\mathcal{O}z=\gamma$, as required. □

If A is any T-matrix at least as efficient as some (C,r), $r>0$, it can be shown (see e.g. [2, p. 213]) that $l^\infty(A)$ inherits nonseparability from l^∞; hence the separability problem raised before is settled for $(\mathcal{O},m)$. In fact, since every Cesàro means (C,r), $r>0$, satisfies just as well the particular summability properties indicated in the following corollary for $\mathcal{O}$, this corollary is therefore a necessary consequence of Theorem 2.

COROLLARY. *Let* $\mathcal{O}$ *denote* $(\mathcal{O},m)$ *for any fixed degree* $m\geqq 1$*. Then*:

(i) $\mathcal{O}$ *has the Borel property*; *in consequence this real T-matrix sums almost all* $z\in\mathscr{H}$ *to their right value* (*their Abel sum*).

(ii) *Any* $z\equiv\{c_k\}\in\mathscr{H}$ *such that the function* $f(s)$ *represented by the Taylor series* $\sum_{k=0}^{\infty}(c_k-c_{k-1})s^k(c_{-1}=0)$ *in* $|s|<1$ *has the boundary point* $s=1$ *as a regular point, or as a point of negative order,*[3] *will be summable* $(\mathcal{O})$ *to its right value* $\lim_{s\to 1^-}f(s)$*, which reduces to* $f(1)$ *if* 1 *is regular.*

(iii) *The divergent Fourier series of any* 2π*-periodic and summable function* $g(x)$*, over* $[-\pi,\pi]$*, is* $\mathcal{O}$*-summable to* $g(x)$*, a.e. on* $[-\pi,\pi]$*; this series is* $\mathcal{O}$*-summable to* $\frac{1}{2}\{g(x+0)+g(x-0)\}$ *if* x *is a jump discontinuity.*

(iv) $(\mathcal{O},m)$ *is consistent with* $(\mathscr{R},r)$ *on* l^∞*, for at least all* $r\in[0,1]$.

(v) $l^\infty(\mathcal{O})$ *retains the nonseparability of* l^∞.

[3] For such a singularity, see Dienes, *Taylor Series*, Oxford, p, 491.

Thus for example, with $z=(0, 0, 1, 0, 0, 1, 0, 0, 1, \cdots)$ whose associated Taylor series, $s^2-s^3+s^5-s^6+s^8-\cdots$, represents $f(s)=s^2/(s^2+s+1)$ on $|s|<1$, which function has 1 as a regular point, the right value for z is $f(1)=1/3$ $(=\lim_{s\to 1^-} f(s))$; hence the $\mathcal{O}$-sum of z is $1/3$. By Corollary 1, the questions raised at the close of §2 are now settled for $(\mathcal{O}, m)$.

The remaining problem, the delicacy of $(\mathcal{O}, m)$, will also be settled by Theorem 2. For since the Cesàro means, as is well known [2, p. 197], is efficient for delicately convergent series (or sequences), it follows from Theorem 2 that the $(\mathcal{O}, m)$ method of any degree $m \geqq 1$ is of the delicate class. In closing, we conjecture that for $m \geqq 2$, $(\mathcal{O}, m)$ is consistent with $(C, 2)$, or in general, for $m \geqq r$, $(\mathcal{O}, m)$ is consistent with (C, r). The case $r=1$ follows from Theorem 2. The first eight of these methods are tabulated below for reference purposes:

$$t_{ik}^{(1)}=4(i+1)/j_{ik}^2;$$

$$t_{ik}^{(2)}=4^2(i+1)^2(i+2)/j_{ik}^4;$$

$$t_{ik}^{(3)}=4^3(i+1)^3(i+2)(i+3)/2j_{ik}^6;$$

$$t_{ik}^{(4)}=4^4(i+1)^4(i+2)^2(i+4)/5j_{ik}^8;$$

$$t_{ik}^{(5)}=4^5(i+1)^5(i+2)^2(i+3)(i+5)/14j_{ik}^{10};$$

$$t_{ik}^{(6)}=4^6(i+1)^6(i+2)^3(i+3)^2(i+4)(i+6)/42(i+5)^2 j_{ik}^{12};$$

$$t_{ik}^{(7)}=4^7(i+1)^7(i+2)^3(i+3)^2(i+4)(i+5)(i+7)/132(i+6)^2 j_{ik}^{14};$$

$$t_{ik}^{(8)}=4^8(i+1)^8(i+2)^4(i+3)^2(i+4)^2(i+5)(i+6)(i+8)/429(i+7)^4 j_{ik}^{16}.$$

REFERENCES

[1] R. G. Cooke, *A new method for summability of divergent sequences*, J. London Math. Soc., 12 (1937), pp 299–304.

[2] _____, *Infinite Matrices and Sequence Spaces*, Macmillan, London, 1950.

[3] N. Kishore, *Rayleigh functions*, Proc. Amer. Math. Soc., (1963), pp. 527–533.

[3a] _____, *A structure for Rayleigh polynomials*, Duke Math. J., 31 (1964), pp. 513–518.

[4] G. N. Watson, *A Treatise on the Theory of Bessel Functions*, Cambridge, 1958. Univ. Press, Cambridge.

[5] R. C. MacCann, *Lower bounds for zeros of Bessel Functions*, Proc. Amer. Math. Soc., 64 (1977), pp 101–102.

SIAM J. MATH. ANAL.
Vol. 17, No. 6, November 1986

018

UNIFORM ASYMPTOTIC EXPANSIONS FOR PROLATE SPHEROIDAL FUNCTIONS WITH LARGE PARAMETERS*

T. M. DUNSTER†

Abstract. By application of the theory for second order linear differential equations with a turning point and a regular (double pole) singularity developed by Boyd and Dunster (this Journal, 17 (1986), pp. 422–450) uniform asymptotic expansions are obtained for prolate spheroidal functions for large γ. The results are uniformly valid for $0 \leqq \mu^2/\gamma^2 \leqq 1+A$ and for $A' \leqq \lambda/\gamma^2 \leqq A''$, where A, A' and A'' are arbitrary real constants such that $0 \leqq A < A' \leqq A'' < \infty$. An asymptotic relationship between λ, μ, γ and the characteristic exponent ν is then derived from the approximations for the spheroidal functions. All the error terms associated with the approximations have explicit bounds given.

Key words. spheroidal wave functions, asymptotic expansions, asymptotic representations

AMS(MOS) subject classifications. Primary 33A55, 34E05, 30E15

1. Introduction. In this paper we shall derive asymptotic approximations to solutions of the spheroidal wave equation, which we express in the form

$$(1.1) \qquad (z^2-1)\frac{d^2v}{dz^2}+2z\frac{dv}{dz}-\left(\lambda+\frac{\mu^2}{(z^2-1)}-\gamma^2(z^2-1)\right)v=0.$$

Throughout this paper we shall restrict the parameters λ, μ, γ to being real and nonnegative. We shall construct asymptotic approximations to solutions of (1.1), for large γ^2, which will be uniformly valid either for $z \equiv x$ real with $0 \leqq x < 1$ (§§2 and 6), or for z complex with $|\arg(z)| \leqq \pi/2$ (§§3, 4 and 6).

There is a fourth parameter associated with (1.1) which is a function of λ, μ, and γ. This parameter is the characteristic exponent and we shall denote it by ν.

When γ is real, solutions of (1.1) are known as prolate spheroidal functions, and in the case where ν and μ are integers they are commonly known as prolate spheroidal wave functions. Prolate spheroidal functions play an important role in many areas of mathematics and physics, and we mention in particular problems of scalar and electromagnetic scattering by a prolate spheroid. In such problems γ is generally a real constant given ab initio, and which is large in the high-frequency case.

There have been a number of approximations for prolate spheroidal wave functions with γ large obtained previously, including the work of Flammer (1957), Müller (1963), Slepian (1965), Streifer (1968), Cloizeaux and Mehta (1972), Miles (1975), and Sink and Eu (1983). These results are essentially heuristic and do not contain error bounds for the difference between the approximate and exact solutions.

The work of Miles (1975) is the most general. He considers four parametric regimes with μ fixed, μ and ν integral, and approximates solutions of (1.1) in both the angular ($0 \leqq x < 1$) and radial ($1 < x < \infty$) cases. His case (ii) is a special case of our results.

* Received by the editors November 5, 1984, and in revised form April 9, 1985. This work was supported by a Science and Engineering Research Council research studentship.

† School of Mathematics, University of Bristol, Bristol BS8 1TW, England. Present address: Department of Mathematical Sciences, The University, Dundee DD1 4HN, Scotland, U.K.

Our results differ from previous results in several significant ways. Our approximations are uniformly valid for μ ranging from 0 to $O(\gamma)$, for large γ^2 and λ, and neither ν nor μ need necessarily be an integer. The approximations either involve Bessel functions or Airy functions, and taken together they are valid over the whole of the right-half plane $|\arg(z)|\leqq\pi/2$. Also, all the approximations in this paper are complete with explicit error bounds.

It is convenient for our analysis to define at this stage new parameters α and β by

$$\alpha=\frac{\mu}{\gamma},\qquad \beta^2=\frac{\lambda}{\gamma^2} \tag{1.2}$$

and a new variable $w(z)$ by

$$w(z)=(z^2-1)^{1/2}v(z). \tag{1.3}$$

Equations (1.2) and (1.3) then transform (1.1) into the differential equation

$$\frac{d^2w}{dz^2}=\left\{\gamma^2\frac{(z^2-z_1^2)(z_2^2-z^2)}{(z^2-1)^2}-\frac{1}{(z^2-1)^2}\right\}w, \tag{1.4}$$

where for convenience we have written

$$z_1=\left[1-\tfrac{1}{2}\left((\beta^4+4\alpha^2)^{1/2}-\beta^2\right)\right]^{1/2}, \tag{1.5a}$$

$$z_2=\left[1+\tfrac{1}{2}\left((\beta^4+4\alpha^2)^{1/2}+\beta^2\right)\right]^{1/2}. \tag{1.5b}$$

In the right-half complex z-plane (1.4) is characterized by a regular singularity at $z=1$ and an irregular singularity at $z=\infty$, and for large γ has turning points at $z=z_1$ and $z=z_2$.

Assume for the moment that β is fixed and positive; we see then from (1.5b) that z_2 is real and lies in the interval $1<(1+\beta^2)^{1/2}\leqq z_2<\infty$ for $0\leqq\alpha^2<\infty$. The character of z_1 as α^2 ranges from 0 to ∞ is determined from (1.5a):

(i) $\alpha^2=0$; z_1 coalesces with the singularity at $z=1$.
(ii) $0<\alpha^2<1+\beta^2$; z_1 is real and lies in the interval $(0,1)$.
(iii) $\alpha^2=1+\beta^2$; z_1 coalesces with the turning point $-z_1$ at $z=0$.
(iv) $\alpha^2>1+\beta^2$; z_1 lies on the imaginary axis.

Our results will be uniformly valid for z_1 real and lying in any closed subinterval of $(0,1]$, and with z_2 bounded away from 1.

We impose then the following restrictions on α and β:

$$0\leqq\alpha^2\leqq 1+A, \tag{1.6a}$$

$$A'\leqq\beta^2\leqq A'', \tag{1.6b}$$

where A, A', and A'' are arbitrary real constants such that

$$0\leqq A<A'\leqq A''<\infty. \tag{1.7}$$

The spheroidal functions we approximate are $\mathrm{Ps}_\nu^{-\mu}(x,\gamma^2)$ and $\mathrm{Qs}_\nu^{-\mu}(x,\gamma^2)$ in the real variable case, and $Ps_\nu^{-\mu}(z,\gamma^2)$, $Qs_\nu^{\mu}(z,\gamma^2)$, $S_\nu^{\mu(3)}(z,\gamma)$ and $S_\nu^{\mu(4)}(z,\gamma)$ in the complex variable case. Definitions of these functions can be found in Arscott (1964, Chap. VIII) and we shall employ Arscott's notation throughout.

We shall frequently refer to the paper by Boyd and Dunster (1986) and so for brevity we shall denote this reference by B&D.

The plan of the paper is as follows. In §§2 and 3 we derive asymptotic expansions as $\gamma \to \infty$ for solutions of (1.1). These expansions are uniformly valid for $0 \leqq x < 1$ in the real variable case, and in certain domains containing the turning point $z = z_1$ and the singularity $z = 1$ in the complex variable case; in both cases the expansions will also be uniformly valid for λ, μ and γ satisfying (1.6). For this we use the theory of B&D which shows that solutions of second order linear differential equations with a turning point and a regular singularity can be approximated by expressions involving Bessel functions of large argument and variable order.

In §4 we derive asymptotic expansions as $\gamma \to \infty$ for solutions of (1.1) which hold at the second turning point $z = z_2$ and at the singularities $z = 1, \infty$. Again we assume (1.6) holds; the expansions are the standard Airy-function approximations for a turning point problem.

In §5 we establish a uniform relationship between ν and λ, μ, γ, as $\gamma \to \infty$, using the results of §3. This is useful insofar as the approximations established in §§2, 3, and 4 for the spheroidal functions involve λ, ν, μ and γ, whereas the spheroidal functions themselves are defined in terms of ν, μ, and γ only. Indeed in the case of the functions Qs and Qs (§6) a relationship between ν and λ, μ, γ is necessary to establish the uniform validity of their approximations.

In §7 we give a summary of the principal results of this paper. This summary is self-contained, in the sense that a reference is given for any term not explicitly defined in the section.

2. A uniform approximation for the spheroidal function $Ps_\nu^{-\mu}(x, \gamma^2)$. In this section we shall construct an asymptotic expansion for large γ for the spheroidal function $Ps_\nu^{-\mu}(x, \gamma^2)$, uniformly valid for $0 \leqq x < 1$ with the parameters λ, μ, γ satisfying the restrictions (1.6). We use the notation of Arscott (1964) for spheroidal functions; thus in terms of associated Legendre functions we have

$$Ps_\nu^{-\mu}(x, \gamma^2) = \sum_{r=-\infty}^{\infty} (-1)^r a_{\nu,r}^{-\mu}(\gamma^2) P_{\nu+2r}^{-\mu}(x),$$

where the coefficients a_r are given by Arscott (1964, p. 168). A second standard solution to (1.1) for $0 \leqq x < 1$ is given by

$$Qs_\nu^{-\mu}(x, \gamma^2) = \sum_{r=-\infty}^{\infty} (-1)^r a_{\nu,r}^{-\mu}(\gamma^2) Q_{\nu+2r}^{-\mu}(x),$$

and we shall establish a uniform asymptotic expansion for this function in §6 (equation (6.13)).

Following B&D (§3) we now apply the following Liouville transformation to (1.4):

$$(2.1) \qquad W(\zeta) = \left(\frac{dx}{d\zeta}\right)^{1/2} w(x), \qquad \left(\frac{d\zeta}{dx}\right)^2 = \left(\frac{4\zeta^2}{\alpha^2 - \zeta}\right) \frac{(x^2 - z_1^2)(z_2^2 - x^2)}{(1 - x^2)^2}.$$

If we take the positive square root of the second of (2.1) and integrate we obtain the following $x - \zeta$ transformation

$$(2.2) \qquad \int_{\alpha^2}^{\zeta} \frac{(\xi - \alpha^2)^{1/2}}{2\xi}\, dt = -\int_{z_1}^{x} \frac{\left[(z_1^2 - t^2)(z_2^2 - t^2)\right]^{1/2}}{1 - t^2}\, dt, \qquad (0 \leqq x \leqq z_1),$$

$$(2.3) \qquad \int_{\alpha^2}^{\zeta} \frac{(\alpha^2 - \xi)^{1/2}}{2\xi}\, d\xi = -\int_{z_1}^{x} \frac{\left[(t^2 - z_1^2)(z_2^2 - t^2)\right]^{1/2}}{1 - t^2}\, dt, \qquad (z_1 \leqq x < 1).$$

We shall denote the value of ζ at $x = 0$ by ζ_0.

Integrating (2.2) and (2.3) gives (see Gradshteyn and Ryzhik (1980, pp. 246, 247, 251))

(2.4)

$$(\zeta-\alpha^2)^{1/2}-\alpha\tan^{-1}\frac{(\zeta-\alpha^2)^{1/2}}{\alpha}=-\frac{(z_2^2-z_1^2)}{z_2}\Pi(\eta,r,t)+z_2E(\eta,t)-x\left(\frac{z_1^2-x^2}{z_2^2-x^2}\right)^{1/2},$$

$$(0\leqq x\leqq z_1),$$

$$(2.5)\qquad (-\zeta+\alpha^2)^{1/2}-\frac{1}{2}\alpha\ln\left(\frac{\alpha+(-\zeta+\alpha^2)^{1/2}}{\alpha-(-\zeta+\alpha^2)^{1/2}}\right)=z_1^2z_2F(\mathscr{X},u)-z_2E(\mathscr{X},u)$$

$$+\frac{1}{x}\{(z_2^2-x^2)(x^2-z_1^2)\}^{1/2}-\frac{z_1^2}{z_2}(z_2^2-1)\Pi(\mathscr{X},s,u),\qquad (z_1\leqq x<1),$$

with

$$(2.6)\qquad r=\frac{z_1^2(1-z_2^2)}{z_2^2(1-z_1^2)},\quad s=\frac{z_2^2-z_1^2}{z_2^2(1-z_1^2)},\quad t=\frac{z_1}{z_2},\quad u=\frac{(z_2^2-z_1^2)^{1/2}}{z_2},$$

$$\eta=\sin^{-1}\left(\frac{x}{z_1}\right)\quad\text{and}\quad\mathscr{X}=\sin^{-1}\left\{\frac{z_2}{x}\left(\frac{x^2-z_1^2}{z_2^2-z_1^2}\right)^{1/2}\right\}.$$

F, E and Π denote elliptic integrals of the first, second and third kind, definitions of which can be found in Gradshteyn and Ryzhik (1980, pp. 904, 905). Let $'$ represent differentiation with respect to ζ; with the above transformations we see that (1.3) is transformed to

$$(2.7)\qquad \frac{d^2W}{d\zeta^2}=\left\{\gamma^2\left(\frac{\alpha^2}{4\zeta^2}-\frac{1}{4\zeta}\right)-\frac{1}{4\zeta^2}+\frac{\psi(\alpha,\zeta)}{\zeta}\right\}W,$$

where

$$(2.8)\qquad \psi(\alpha,\zeta)=\dot{x}^{1/2}\frac{d^2}{d\zeta^2}(\dot{x}^{-1/2})+\frac{1}{4\zeta^2}-\left(\frac{\dot{x}}{1-x^2}\right)^2.$$

On using the second of (2.1) ψ can be expressed explicitly as

$$(2.9)\quad \psi(\alpha,\zeta)=\frac{\zeta+4\alpha^2}{16(\zeta-\alpha^2)^2}+\frac{(x^2-1)(\zeta-\alpha^2)}{16\zeta}$$

$$\cdot\left[\frac{5(4\alpha^2+\beta^4)x^2(x^2-1)}{\left(\alpha^2+\beta^2(x^2-1)-(x^2-1)^2\right)^3}-\frac{2(3\beta^2+2)x^2-2\beta^2-4}{\left(\alpha^2+\beta^2(x^2-1)-(x^2-1)^2\right)^2}\right].$$

Since β is essentially fixed (see (1.6b)) we shall for simplicity suppress any dependence of β in all functions. Thus in (2.9) $\psi(\alpha,\zeta)\equiv\psi(\alpha,\beta,\zeta)$. Before we can identify the solutions of (1.4) with those of (2.7) it is necessary for us to determine the asymptotic behavior of $\zeta(x)$ as $x\to1-$. This is achieved by comparing the limiting behavior of

both sides of (2.3) as $x\to 1-$ and $\zeta\to 0+$. First we make the splitting

$$-\int_{z_1}^{x}\frac{\left[\left(t^2-z_1^2\right)\left(z_2^2-t^2\right)\right]^{1/2}}{1-t^2}\,dt=-\int_{z_1}^{x}\frac{\left[\left(t^2-z_1^2\right)\left(z_2^2-t^2\right)\right]^{1/2}-\alpha}{1-t^2}\,dt-\alpha\int_{z_1}^{x}\frac{dt}{1-t^2}$$

and on observing that the first integral on the right-hand side is bounded as $x\to 1-$ we find that

$$1-x=a_1(\alpha)\zeta+O(\zeta^2),\qquad(\zeta\to 0+),\tag{2.10}$$

where

$$a_1(\alpha)=\frac{1}{2}\left(\frac{1-z_1}{\alpha^2}\right)(1+z_1)^{-1}e^{2+\rho_1(\alpha)},\tag{2.11}$$

with

$$\rho_1(\alpha)=\frac{2}{\alpha}\int_{z_1}^{1}\frac{\left[\left(t^2-z_1^2\right)\left(z_2^2-t^2\right)\right]^{1/2}-\alpha}{1-t^2}\,dt.\tag{2.12}$$

Note that (2.10) is uniformly valid for all values of α under consideration; when $\alpha=0$ the limiting value of (2.10), namely $1-x\sim(1/2\beta)\zeta$, applies.

From Theorem 1 of B&D, with u replaced by γ, we obtain the following solution of (2.7):

$$W_{2n+1,1}(\gamma,\alpha,\zeta)=\zeta^{1/2}J_{\gamma\alpha}(\gamma\zeta^{1/2})\sum_{s=0}^{n}\frac{A_s(\alpha,\zeta)}{\gamma^{2s}}+\frac{\zeta}{\gamma}J'_{\gamma\alpha}(\gamma\zeta^{1/2})\sum_{s=0}^{n-1}\frac{B_s(\alpha,\zeta)}{\gamma^{2s}}+\varepsilon_{2n+1,1}(\gamma,\alpha,\zeta).\tag{2.13}$$

The functions $A_s(\alpha,\zeta)$ and $B_s(\alpha,\zeta)$ are defined recursively for $s=0,1,2,\cdots$ from (2.9a, b) of B&D with $A_0(\alpha,\zeta)=1$. Bounds for the error function ε and its derivative are given by (3.8) of B&D: these bounds are somewhat complicated but their significance is explained in §3 of B&D. Two important consequences are (i) $\varepsilon_{2n+1,1}(\zeta)[\zeta^{1/2}J_{\gamma\alpha}(\gamma\zeta^{1/2})]^{-1}$ is $O(\gamma^{-2n-1})$ for large γ, uniformly with respect to ζ in the interval $(0,\zeta_0]$ and with respect to α^2 in the interval $[0,1+A]$; and (ii) $\varepsilon_{2n+1,1}(\zeta)=\zeta^{1/2}J_{\gamma\alpha}(\gamma\zeta^{1/2})O(\zeta)$ as $\zeta\to 0$. From the latter we see at once that $W_{2n+1,1}$ is recessive at $\zeta=0$ and hence is a multiple of a solution of (1.1) that is recessive at $x=1$. This solution is the spheroidal function $\mathrm{Ps}_\nu^{-\mu}(x,\gamma^2)$. From (1.1), (1.3), and (2.1) we therefore deduce that

$$\mathrm{Ps}_\nu^{-\mu}(x,\gamma^2)=C_{2n+1,1}\left(\frac{\alpha^2-\zeta}{\alpha^2-\beta^2(1-x^2)-(1-x^2)^2}\right)^{1/4}\zeta^{-1/2}W_{2n+1,1}(\gamma,\alpha,\zeta),\qquad(0\leqq x<1),\tag{2.14}$$

where the coefficient $C_{2n+1,1}$ is independent of x. To determine $C_{2n+1,1}$ we evaluate the limiting value of the ratio of the two sides of (2.14) as $x\to 1-$ and $\zeta\to 0+$. Firstly we have, from Arscott (1964, p. 171)

$$\mathrm{Ps}_\nu^{-\mu}(x,\gamma^2)=\frac{A_\nu^{-\mu}(\gamma^2)}{2^{\mu/2}\Gamma(1+\mu)}(1-x)^{\mu/2}\{1+O(1-x)\},\qquad(x\to 1-),\tag{2.15}$$

where

$$A_\nu^{-\mu}(\gamma^2)=\sum_{r=-\infty}^{\infty}(-1)^r a_{\nu,r}^{-\mu}(\gamma^2).$$

From (2.13) we see that the right-hand side of (2.14) is equal to

(2.16)

$$C_{2n+1,1}\left(\frac{\gamma}{2}\right)\frac{\zeta^{\mu/2}}{\Gamma(1+\mu)}\left(1+\alpha\sum_{s=0}^{n-1}\frac{B_s(\alpha,0)}{\gamma^{2s+1}}+\sum_{s=1}^{n}\frac{A_s(\alpha,0)}{\gamma^{2s}}\right)(1+O(\zeta))\quad\text{as }\zeta\to 0+.$$

(In (2.16) we have used the fact that $J_\nu(x)\sim(\frac{1}{2}x)^\nu/\Gamma(\nu+1)$ as $x\to 0$, (see, for example, Olver (1974, p. 436)).)

If we divide (2.16) by (2.15) and use (2.10) we obtain the following exact expression for the coefficient $C_{2n+1,1}$:

$$A_\nu^{-\mu}(\gamma^2)\left(\frac{2a_1(\alpha)}{\gamma^2}\right)^{\mu/2}\left(1+\alpha\sum_{s=0}^{n-1}\frac{B_s(\alpha,0)}{\gamma^{2s+1}}+\sum_{s=1}^{n}\frac{A_s(\alpha,0)}{\gamma^{2s}}\right)^{-1}. \tag{2.17}$$

We note here that there is a certain degree of arbitrariness in (2.17), insofar as there is an arbitrary integration constant λ_s associated with each of $A_s(\alpha,0)$ $(s=1,2,\cdots)$ (cf. B&D, eq. (2.9b)). A natural choice is choose them so that $A_s(\alpha,0)=0$ $(s=1,2,\cdots)$, thus removing the second series in (2.17).

3. Uniform approximations for spheroidal functions in domains containing the turning point $z=z_1$. In this section we shall construct asymptotic expansions for large γ for solutions of (1.1) with complex argument (denoted by z). The results will be uniformly valid with respect to z in certain domains containing the turning point $z=z_1$ and with parameters satisfying the restrictions of (1.6). The standard solutions we approximate in this section are $Ps_\nu^{-\mu}(z,\gamma^2)$, $S_\nu^{\mu(3)}(z,\gamma)$ and $S_\nu^{\mu(4)}(z,\gamma)$. Our notation is that of Arscott (1964, Chap. VIII, pp. 169, 170, 174); thus in terms of associated Legendre functions we have

$$Ps_\nu^{-\mu}(z,\gamma^2)=\sum_{r=-\infty}^{\infty}(-1)^r a_{\nu,r}^{-\mu}(\gamma^2)P_{\nu+2r}^{-\mu}(z),$$

and this function is recessive at $z=1$. For the S-functions we have

$$S_\nu^{\mu(j)}(z,\gamma)=\frac{(z^2-1)^{-\mu/2}z^\mu}{A_\nu^\mu(\gamma^2)}\sum_{r=-\infty}^{\infty}a_{\nu,r}^{\mu}(\gamma^2)\psi_{\nu+2r}^{(j)}(\gamma z),\qquad (j=3,4)$$

where $\psi^{(j)}$ are spherical Bessel functions. $S_\nu^{\mu(3)}$ is recessive as $z\to\infty$ for $0<\arg(z)<\pi$, and $S_\nu^{\mu(4)}$ is recessive as $z\to\infty$ for $-\pi<\arg(z)<0$.

In §6 we shall derive uniform asymptotic expansions for the spheroidal functions $Qs_\nu^\mu(z,\gamma^2)$ and $Qs_\nu^{-\mu}(x,\gamma^2)$. For brevity we shall only consider the range $|\arg(z)|\leqq\pi/2$; appropriate connection formulae for spheroidal functions can be used for other ranges of $\arg(z)$.

We use the theory of §5 of B&D. The appropriate Liouville transformation is again given by (2.1), with x replaced by z and ζ now considered as a complex variable. Integration of the second of (2.1) now yields, in place of (2.2) and (2.3), the following

z-ζ transformation

$$(3.1)\quad \int_{\alpha^2}^{\zeta} \frac{(\alpha^2-\xi)^{1/2}}{2\xi}\,d\xi = -\int_{z_1}^{x} \frac{\left[(t^2-z_1^2)(z_2^2-t^2)\right]^{1/2}}{1-t^2}\,dt, \qquad z\in\Delta,\quad \zeta\in\underline{\Delta},$$

where Δ is a certain subdomain of $|\arg(z)|\leqq\pi/2$ and $\underline{\Delta}$ is the corresponding ζ map of this domain. Equation (3.1) as it stands does not give a 1-1 mapping between z and ζ as there is a singularity in the transformation at $z=z_2$. Therefore we shall now choose a branch cut in $\underline{\Delta}$ from $\zeta=\zeta_2$ to $\zeta=-\infty$ along the negative real ζ-axis, where ζ_2 denotes the map given by (3.1) of z at $z=z_2$. With this cut we take the integrands of both sides of (3.1) to be negative when z and ζ lie in the intervals $(1,z_2)$ and $(\zeta_2,0)$ respectively, and we take them to be continuous elsewhere in Δ and $\underline{\Delta}$. Equation (3.1) now gives a continuous 1-1 mapping between the independent variables z and ζ. Domains Δ and $\underline{\Delta}$ are shown in Figs. 1a and 2; in Fig. 1a, Γ_1 and Γ_2 are conjugate curves given by

$$(3.2)\quad \operatorname{Im}\int_{z_2}^{z} \frac{\left[(t^2-z_1^2)(z_2^2-t^2)\right]^{1/2}}{1-t^2}\,dt=0.$$

A knowledge of the general configuration of the *level curves* in the ζ-plane, defined by

$$\operatorname{Im}\Phi_\mu^{(j)}(\gamma\zeta^{1/2}) \equiv \operatorname{Re}\int_{\hat{X}_\mu^2}^{\gamma^2\zeta} \frac{\left(\hat{X}_\mu^2-t\right)^{1/2}}{2t}\,dt = \text{constant},$$

is necessary in applying the theory of §5 of B&D. (A definition of $\hat{X}_\mu$ is given in the summary (equation (7.13)).) These curves are illustrated in Fig. 2. The corresponding configuration of these curves in the ζ-plane depends mainly on whether ζ_2 lies inside or outside the pear-shaped region $S_\alpha^{(0)}$ bounded by the level curve $\operatorname{Im}\Phi_\mu^{(0)}(\gamma\zeta^{1/2})=0$, or whether ζ_2 lies on this level curve. The three cases are illustrated in Figs. 1b, c, d.

It is straightforward to show that the curve $\operatorname{Im}\Phi_\mu^{(0)}(\gamma\zeta^{1/2})=0$ intersects the negative ζ-axis at $\zeta=-(1/\gamma^2)\hat{X}_\mu^2(\tau^2-1)$, where τ is the positive root of

$$(3.3a)\quad \left(\frac{\tau+1}{\tau-1}\right)-e^{2\tau}=0, \qquad (\tau=1.19968\cdots).$$

It follows then that ζ_2 will lie outside $S_\alpha^{(0)}$ iff

$$(3.3b)\quad |\zeta_2| > \frac{1}{\gamma^2}\hat{X}_\mu^2(\tau^2-1).$$

With the transformation (3.1) we obtain (2.7) with ζ a complex variable. Before we obtain solutions to this equation it is necessary for us to determine the asymptotic behavior of z, regarded as a function of ζ, as $|\zeta|\to\infty$. It is easy to see from (3.1) that

$$(3.4a)\quad z\sim \mp i|\zeta|^{1/2}, \qquad (\zeta\to-\infty\pm i0),$$

where $\pm i0$ denotes ζ above or below the cut running from $\zeta=\zeta_2$ to $\zeta=-\infty$. We will, in fact, need the next term in this asymptotic relationship, and this is not so straightforward to deduce.

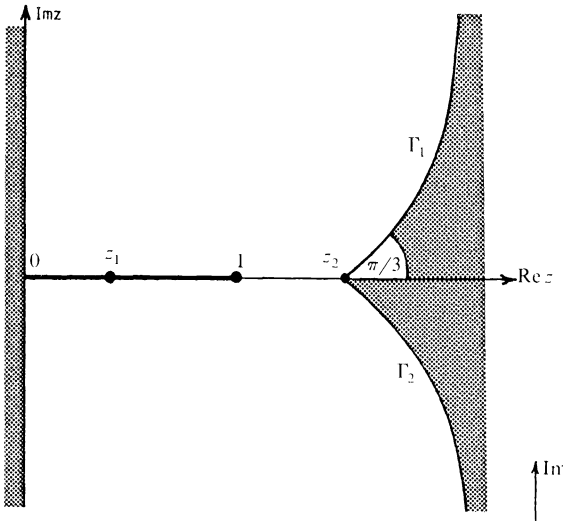


FIG. 1a. *Domain* Δ.

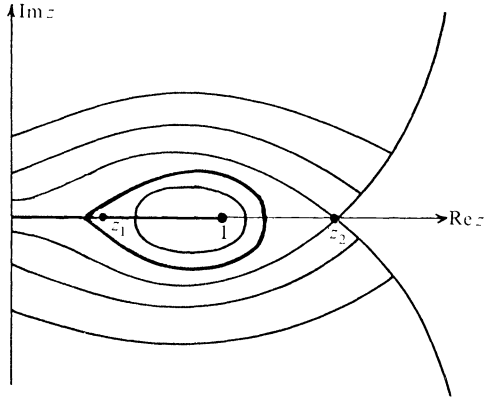


FIG. 1b. *Level curves for* $|\zeta_2|>\gamma^{-2}\hat{X}_\mu^2(\tau^2-1)$.

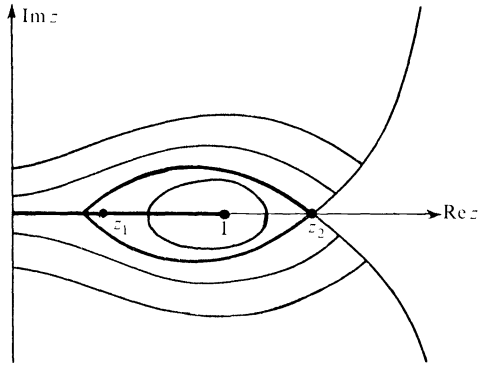


FIG 1c. *Level curves for* $|\zeta_2|=\gamma^{-2}\hat{X}_\mu^2(\tau^2-1)$.

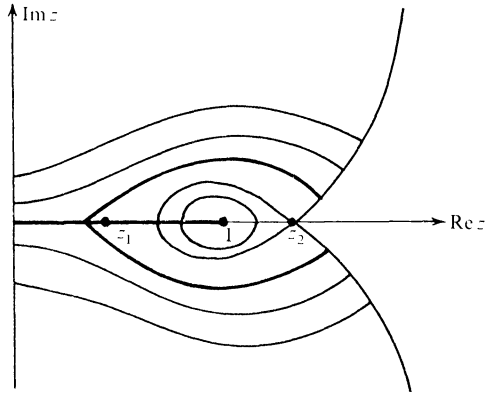


FIG. 1d. *Level curves for* $|\zeta_2|<\gamma^{-2}\hat{X}_\mu^2(\tau^2-1)$.

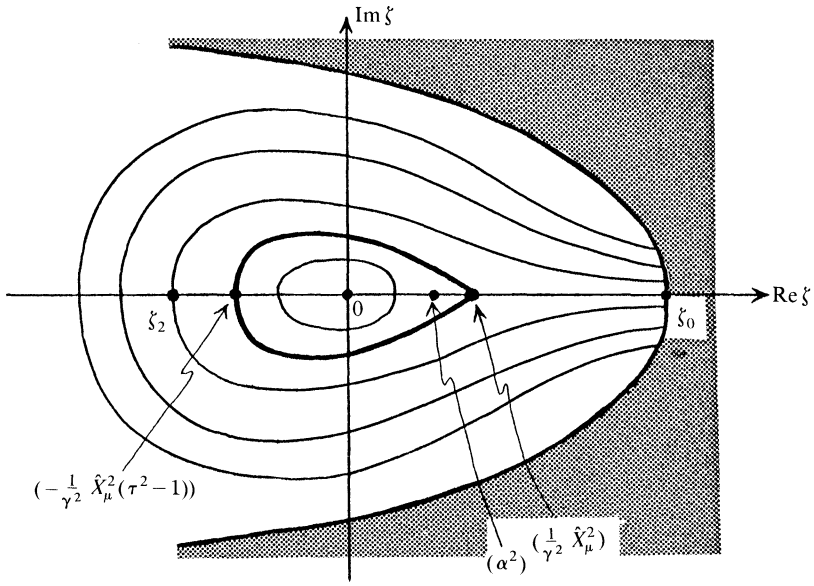


FIG. 2. *Domain* Δ *with level curves* $\operatorname{Im}\Phi_\mu^{(j)}(\gamma\zeta^{1/2})=\text{constant}$.

We first determine the asymptotic behavior of $z(\zeta)$ as $\zeta\to-\infty-i0$. If we write $\zeta=-R^2-i0$, where R is real and positive, then from (3.4a) we can write the corresponding value of $z(\zeta)$ as

$$z=iR+\tilde{z}(R) \tag{3.4b}$$

where

$$\frac{\tilde{z}(R)}{R}=o(1) \quad \text{as } R\to\infty, \tag{3.4c}$$

We shall now determine an asymptotic expression for $\tilde{z}$ as $R\to\infty$. To do so we consider the integral

$$J(R)=\int_0^{iR+\tilde{z}}\frac{\left[\left(t^2-z_1^2\right)\left(z_2^2-t^2\right)\right]^{1/2}+it^2}{1-t^2}\,dt \tag{3.5}$$

where the path of integration is shown in Fig. 3 and the branches of the square roots are the same as for (2.3). Now if we make the splitting of (3.5) as

$$J(R)=i\int_0^{iR+\tilde{z}}\frac{t^2}{1-t^2}\,dt+\int_0^{z_2}\frac{\left[\left(t^2-z_1^2\right)\left(z_2^2-t^2\right)\right]^{1/2}}{1-t^2}\,dt$$

$$+\int_{z_2}^{iR+\tilde{z}}\frac{\left[\left(t^2-z_1^2\right)\left(z_2^2-t^2\right)\right]^{1/2}}{1-t^2}\,dt$$

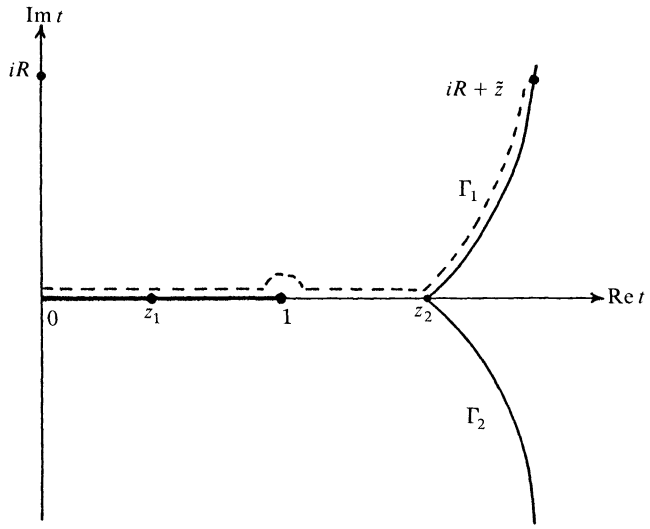


FIG. 3. *Path of integration* ----.

and equate the imaginary parts of both sides, we find on employing (3.2) that

$$\text{(3.6)}\qquad \operatorname{Im}\{J(R)\} = \operatorname{Re}\int_0^{iR+\tilde{z}} \frac{t^2}{1-t^2}\,dt + \operatorname{Im}\int_0^{z_2} \frac{\left[\left(t^2-z_1^2\right)\left(z_2^2-t^2\right)\right]^{1/2}}{1-t^2}\,dt$$

$$= -\tilde{z}(R) + \int_0^{z_1} \frac{\left[\left(z_1^2-t^2\right)\left(z_2^2-t^2\right)\right]^{1/2}}{1-t^2}\,dt + \frac{\alpha\pi}{2}.$$

We next show that the left-hand side of (3.6) is $o(1)$ as $R\to\infty$; we do this by observing that the integrand of (3.5) is $O(t^{-2})$ as $t\to\infty$ between the imaginary axis and Γ_1, and so from Cauchy's theorem we can deform the contour of integration of J to lie along the imaginary axis as $R\to\infty$. Thus from (3.4c) and (3.5) we have

$$J(R) = \int_0^{iR} \frac{\left[\left(t^2-z_1^2\right)\left(z_2^2-t^2\right)\right]^{1/2} + it^2}{1-t^2}\,dt + o(1), \qquad (R\to\infty).$$

Since the integral of this equation is real we deduce that $\operatorname{Im}\{J(R)\} = o(1)$ as $R\to\infty$ and therefore from (3.6) we have the desired expression for $\tilde{z}$:

$$\text{(3.7a)}\qquad \tilde{z}(R) = \sigma + \frac{\alpha\pi}{2} + o(1), \qquad (R\to\infty),$$

where

$$\text{(3.7b)}\qquad \sigma = \int_0^{z_1} \frac{\left[\left(z_1^2-t^2\right)\left(z_2^2-t^2\right)\right]^{1/2}}{1-t^2}\,dt = \int_{\alpha^2}^{\zeta_0} \frac{(t-\alpha^2)^{1/2}}{2t}\,dt.$$

The same argument holds when $\zeta\to-\infty+i0$, and so from (3.4b) and (3.7a) we have

the following asymptotic behavior of $z(\zeta)$ as $|\zeta|\to\infty$:

$$z=\mp i|\zeta|^{1/2}+\sigma+\frac{\alpha\pi}{2}+o(1),\qquad(\zeta\to-\infty\pm i0). \tag{3.8}$$

An important consequence of (3.8), on using (2.9), is that the ζ-derivatives

$$\psi^{(s)}(\alpha,\zeta)=O\big(|\zeta|^{-s-1}\big),\qquad(\zeta\to-\infty\pm i0). \tag{3.9}$$

We now apply Theorem 3 of B&D, with u replaced by γ, to obtain the following pair of linearly independent solutions of (2.7) holomorphic in $\underline{\Delta}$:

(3.10)

$$W^{(j)}_{2n+1}(\gamma,\alpha,\zeta)=\zeta^{1/2}\mathscr{C}^{(j)}_{\gamma\alpha}(\gamma\zeta^{1/2})\sum_{s=0}^{n}\frac{A_s(\alpha,\zeta)}{\gamma^{2s}}$$
$$+\frac{\zeta}{\gamma}\mathscr{C}^{(j)'}_{\gamma\alpha}(\gamma\zeta^{1/2})\sum_{s=0}^{n-1}\frac{B_s(\alpha,\zeta)}{\gamma^{2s}}+\varepsilon^{(j)}_{2n+1}(\gamma,\alpha,\zeta),\qquad(j=0,1).$$

In (3.10) n is a nonnegative integer, and $\mathscr{C}^{(j)}_{\gamma\alpha}(\gamma\zeta^{1/2})$ denotes $2J_{\gamma\alpha}(\gamma\zeta^{1/2})$, $H^{(1)}_{\gamma\alpha}(\gamma\zeta^{1/2})$ for $j=0,1$ respectively. We define $\zeta^{1/2}$ here such that $0\leqq\arg(\zeta^{1/2})<\pi$.

For both solutions we associate a reference point $\tilde{\zeta}^{(j)}$ in $\underline{\Delta}$ from which error bounds on $\varepsilon^{(j)}_{2n+1}$ are determined. Since $W^{(0)}_{2n+1}$ is recessive at $\zeta=0$ we choose $\tilde{\zeta}^{(0)}=0$. For the other solution $W^{(1)}_{2n+1}$ we shall choose two reference points: $\tilde{\zeta}^{(1)}=-\infty+i0$ and $\tilde{\zeta}^{(1)}=-\infty-i0$. We shall denote $W^{(1)}_{2n+1}$ with reference point $\tilde{\zeta}^{(1)}=-\infty+i0$ by W^{+}_{2n+1} and $W^{(1)}_{2n+1}$ with reference point $\tilde{\zeta}^{(1)}=-\infty-i0$ by W^{-}_{2n+1}. Both W^{+}_{2n+1} and W^{-}_{2n+1} are expressible in the form (3.10) with $\mathscr{C}^{(1)}_{\gamma\alpha}\equiv H^{(1)}_{\gamma\alpha}$, and both are holomorphic in $\underline{\Delta}$, but the error terms (which we denote by ε^{+}_{2n+1} and ε^{-}_{2n+1}) are different in the two cases.

The error functions in (3.10) satisfy the following bounds:

(3.11a)

$$\frac{|\varepsilon^{(0)}_{2n+1}(\gamma,\alpha,\zeta)|}{\zeta^{1/2}M^{(2)}_{\gamma\alpha}(\gamma\zeta^{1/2})},\ \frac{|\partial\varepsilon^{(0)}_{2n+1}(\gamma,\alpha,\zeta)/\partial\zeta|}{\frac{1}{2}\gamma N^{(2)}_{\gamma\alpha}(\gamma\zeta^{1/2})+\frac{1}{2}\zeta^{-1/2}M^{(2)}_{\gamma\alpha}(\gamma\zeta^{1/2})}$$
$$\leqq\frac{1}{\gamma^{2n+1}}\kappa E^{(0)}_{\gamma\alpha}(\gamma\zeta^{1/2})^{-1}\mathscr{V}_{0,\zeta}\big\{(\xi-\alpha^2)^{1/2}B_n(\xi)\big\}\exp\Big(\frac{\kappa}{\gamma}\mathscr{V}_{0,\zeta}\big\{(\xi-\alpha^2)^{1/2}B_0(\xi)\big\}\Big),$$

when $\zeta\in\underline{\Delta}$;

(3.11b)

$$\frac{|\varepsilon^{\pm}_{2n+1}(\gamma,\alpha,\zeta)|}{\zeta^{1/2}M^{(2)}_{\gamma\alpha}(\gamma\zeta^{1/2})},\ \frac{|\partial\varepsilon^{\pm}_{2n+1}(\gamma,\alpha,\zeta)/\partial\zeta|}{\frac{1}{2}\gamma N^{(2)}_{\gamma\alpha}(\gamma\zeta^{1/2})+\frac{1}{2}\zeta^{-1/2}M^{(2)}_{\gamma\alpha}(\gamma\zeta^{1/2})}$$
$$\leqq\frac{1}{\gamma^{2n+1}}\kappa E^{(1)}_{\gamma\alpha}(\gamma\zeta^{1/2})^{-1}\mathscr{V}_{\zeta,-\infty\pm i0}\big\{(\xi-\alpha^2)^{1/2}B_n(\xi)\big\}$$
$$\times\exp\Big(\frac{\kappa}{\gamma}\mathscr{V}_{\zeta,-\infty\pm i0}\big\{(\xi-\alpha^2)^{1/2}B_0(\xi)\big\}\Big),$$

when $\zeta\in\underline{\Delta}^{\pm}$ (see Figs. 4b and 5b).

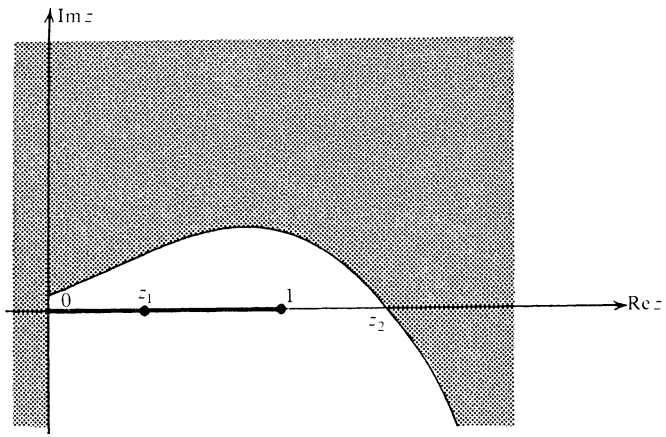


FIG. 4a. *Domain* Δ^+ *with* $|\zeta_2| > \gamma^{-2}\hat{X}_\mu^2(\tau^2-1)$. (*The domain* Δ^- *is the conjugate of* Δ^+.)

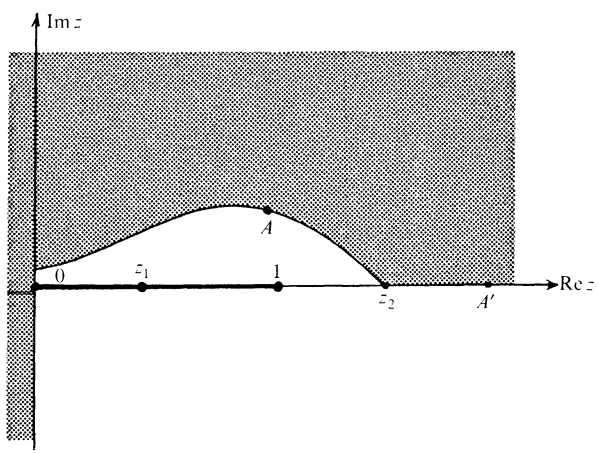


FIG. 4c. *Domain* Δ^+; *maximal region of validity.*

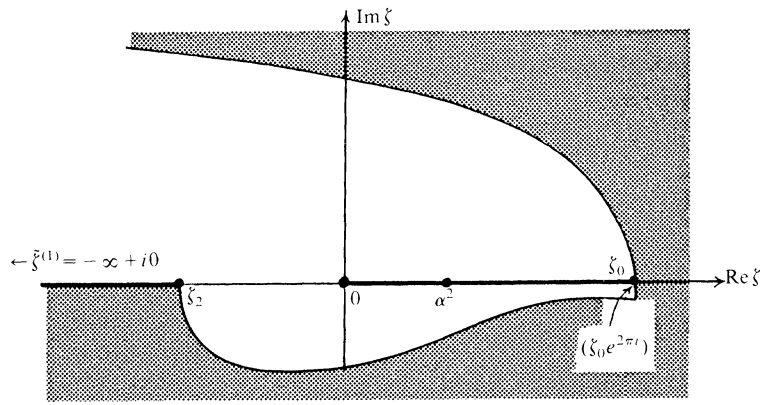


FIG. 4b. *Domain* $\underline{\Delta}^+$ *with* $|\zeta_2| > \gamma^{-2}\hat{X}_\mu^2(\tau^2-1)$. (*The domain* $\underline{\Delta}^-$ *is the conjugate of* $\underline{\Delta}^+$.)

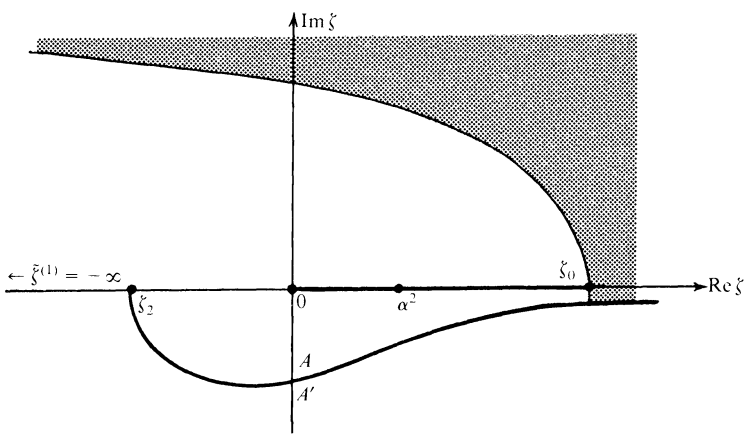


FIG. 4d. *Domain* $\underline{\Delta}^+$ *with branch cut along level curve.*

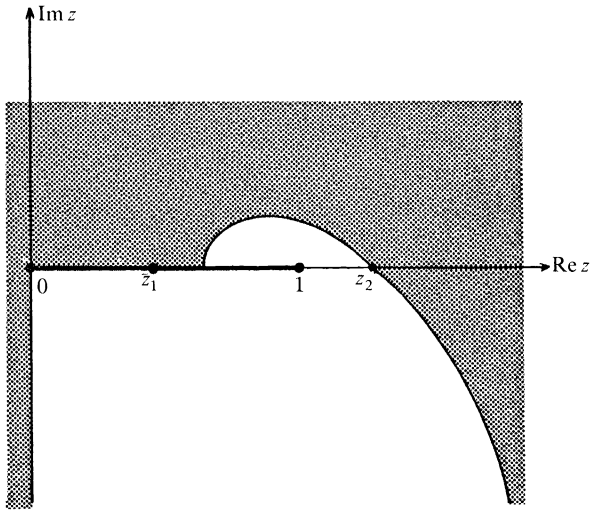


FIG. 5a. *Domain* Δ^+ *with* $|\zeta_2|<\gamma^{-2}\hat{X}_\mu^2(\tau^2-1)$.

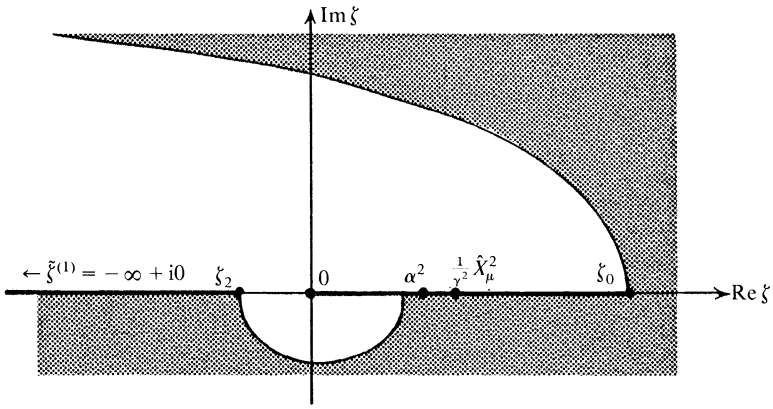


FIG. 5b. *Domain* $\underline{\Delta}^+$ *with* $|\zeta_2|<\gamma^{-2}\hat{X}_\mu^2(\tau^2-1)$.

The weight and modulus functions E, M and N, and the constant κ are defined as in §5 of B&D. All variations in (3.11a, b) are taken on progressive paths (cf. B&D, (5.14)), and their existence is a consequence of (3.9) (cf. Olver (1974, p. 445)).

We now proceed to identify the solutions (3.10) with standard spheroidal functions. Firstly $(d\zeta/dz)^{1/2}(z^2-1)^{1/2}Ps_\nu^{-\mu}(z,\gamma^2)$, considered as a function ζ, and $W_{2n+1}^{(0)}(\gamma,\alpha,\zeta)$ are both recessive at $\zeta=0$ and both are solutions of (2.7). It follows immediately that there exists a coefficient $C_{2n+1}^{(0)}$, independent of z, such that

$$(3.12)\quad Ps_\nu^{-\mu}(z,\gamma^2)=C_{2n+1}^{(0)}\left(\frac{\alpha^2-\zeta}{\alpha^2+\beta^2(z^2-1)-(z^2-1)^2}\right)^{1/4}\zeta^{-1/2}W_{2n+1}^{(0)}(\gamma,\alpha,\zeta).$$

We know that (Arscott (1964, p. 171))

$$(3.13)\qquad Ps_\nu^{-\mu}(z,\gamma^2)=\frac{A_\nu^{-\mu}(\gamma^2)(z-1)^{\mu/2}}{2^{\mu/2}\Gamma(1+\mu)}\left(1+O(|z-1|)\right),\qquad (z\to 1),$$

and so from the same method of determining (2.17) we find from (3.10), (3.12) and (3.13) that

$$C_{2n+1}^{(0)}=\frac{1}{2}C_{2n+1,1}. \tag{3.14}$$

Next, we observe that $(d\zeta/dz)^{1/2}(z^2-1)^{1/2}S_\nu^{\mu(3)}(z,\gamma)$, considered as a function of ζ, and $W_{2n+1}^-(\gamma,\alpha,\zeta)$ are both recessive at $\zeta=\infty$ in $\underline{\Delta}^-$ and are both solutions of (2.7). Therefore we have the identity

$$S_\nu^{\mu(3)}(z,\gamma)=C_{2n+1}^-\left(\frac{\alpha^2-\zeta}{\alpha^2+\beta^2(z^2-1)-(z^2-1)^2}\right)^{1/4}\zeta^{-1/2}W_{2n+1}^-(\gamma,\alpha,\zeta) \tag{3.15}$$

where the coefficient C_{2n+1}^- is independent of ζ.

We can conveniently determine this constant by comparing both sides of (3.15) as $\zeta\to\infty$ in $\underline{\Delta}^-$ and $z\to\infty$ in Δ^- (see Fig. 4a). For the left-hand side we have (Meixner and Schäfke (1954, p. 293))

$$S_\nu^{\mu(3)}(z,\gamma)=\frac{e^{i(\gamma z-\pi(\nu+1)/2)}}{\gamma z}\left(1+O\left(\frac{1}{z}\right)\right),\qquad (z\to\infty). \tag{3.16}$$

For the right-hand side we have the following asymptotic behavior (on using the asymptotic expression for $H_\mu^{(1)}(\gamma\zeta^{1/2})$ with large argument (Abramowitz and Stegun (1965, p. 364)):

$$C_{2n+1}^-\left(\frac{2}{\pi\gamma}\right)^{1/2}\frac{1}{|z|}e^{i(\gamma\zeta^{1/2}-\mu\pi/2-3\pi/4)}\left\{\sum_{s=0}^{n}\frac{A_s(-\infty)}{\gamma^{2s}}+\frac{1}{\gamma}\sum_{s=0}^{n-1}\frac{\tilde{B}_s(-\infty)}{\gamma^{2s}}+o(1)\right\}, \tag{3.17}$$

$$(\zeta\to-\infty-i0).$$

In (3.17) $\tilde{B}_s(-\infty)\equiv\lim_{\zeta\to-\infty}\{(-\zeta)^{1/2}B_s(\zeta)\}$. By comparing (3.16) and (3.17), and invoking (3.8) we obtain the following exact expression for C_{2n+1}^-:

$$ie^{i\chi}\left(\frac{\pi}{2\gamma}\right)^{1/2}e^{\mu\pi i/2}\left\{\sum_{s=0}^{n}\frac{A_s(-\infty)}{\gamma^{2s}}+\frac{1}{\gamma}\sum_{s=0}^{n-1}\frac{\tilde{B}_s(-\infty)}{\gamma^{2s}}\right\}^{-1} \tag{3.18}$$

where

$$\chi=\frac{\mu\pi}{2}-\frac{\nu\pi}{2}+\gamma\sigma-\frac{3\pi}{4}. \tag{3.19}$$

By the same procedure we find that

$$S_\nu^{\mu(4)}(z,\gamma)=C_{2n+1}^+\left(\frac{\alpha^2-\zeta}{\alpha^2+\beta^2(z^2-1)-(z^2-1)^2}\right)^{1/4}\zeta^{-1/2}W_{2n+1}^+(\gamma,\alpha,\zeta), \tag{3.20}$$

where

$$(3.21)\qquad C_{2n+1}^{+}=ie^{-i\chi}\left(\frac{\pi}{2\gamma}\right)^{1/2}e^{\mu\pi i/2}\left\{\sum_{s=0}^{n}\frac{A_s(-\infty)}{\gamma^{2s}}+\frac{1}{\gamma}\sum_{s=0}^{n-1}\frac{\tilde{B}_s(-\infty)}{\gamma^{2s}}\right\}^{-1}.$$

We finally remark that Δ, Δ^- and Δ^+ are not the maximal regions of validity for our uniform expansions for $Ps_\nu^{-\mu}(z,\gamma^2)$, $S_\nu^{\mu(3)}(z,\gamma)$ and $S_\nu^{\mu(4)}(z,\gamma)$ respectively; if instead of choosing a branch cut along the negative ζ-axis from $\zeta=\zeta_2$ to $\zeta=-\infty$ for the $z-\zeta$ transformation (3.1) we could for instance have chosen the cut to be along the level curve passing through ζ_2 (see Fig. 4d). The extended region of validity for the asymptotic expansion for $S_\nu^{\mu(4)}$ in this case is shown in Fig. 4c. For the function $Ps_\nu^{-\mu}$ we could extend the region of the validity Δ by choosing a cut in the z-plane along the real axis from $z=z_2$ to $z=\infty$; the region of validity Δ in this case would be extended to all points in the right-half plane $|\arg(z)|\leqq\pi/2$ except points on or near this cut. The corresponding ζ-domain $\underline{\Delta}$ would in this case be a Riemann sheet.

However, if one is considering uniform expansions for any pair of $Ps_\nu^{-\mu}$, $S_\nu^{\mu(3)}$ and $S_\nu^{\mu(4)}$ in tandem the branch cut from $\zeta=\zeta_2$ to $\zeta=-\infty$ for the $z-\zeta$ transformation would seem to be the most convenient choice. Indeed this is our choice throughout this paper.

4. Uniform approximations for spheroidal functions in a domain containing the turning point $z=z_2$. In this section we establish asymptotic expansions that hold in a domain containing the second turning point $z=z_2$. The expansions are the standard Airy function approximations for a turning point problem, the theory of which is concisely presented in Olver [1974, Chap. 11].[1] As in §3 the standard solutions we consider are $Ps_\nu^{-\mu}(z,\gamma^2)$, $S_\nu^{\mu(3)}(z,\gamma)$, and $S_\nu^{\mu(4)}(z,\gamma)$. We restrict the parameters λ, μ and γ to satisfy (1.6) and we only consider values of z such that $|\arg(z)|\leqq\pi/2$. To avoid a clash of notation with §3 we shall use the notation of Olver but with each term written with a circumflex (ˆ).

From (3.02) of Olver we see that the appropriate Liouville transformation is

$$(4.1)\qquad \hat{W}(\hat{\zeta})=\left(\frac{d\hat{\zeta}}{dz}\right)^{1/2}w(z),\qquad \frac{2}{3}\hat{\zeta}^{3/2}=-\int_{z_2}^{z}\frac{\left[(t^2-z_1^2)(z_2^2-t^2)\right]^{1/2}}{t^2-1}\,dt,$$

which throws (1.4) into the form

$$(4.2)\qquad \frac{d^2\hat{W}}{d\hat{\zeta}^2}=\{\gamma^2\hat{\zeta}+\hat{\psi}(\hat{\zeta})\}\hat{W},$$

where

$$(4.3)\quad \hat{\psi}(\hat{\zeta})=\frac{5}{16\hat{\zeta}^2}-\frac{\hat{\zeta}(z^2-1)}{4}\times\left[\frac{5(4\alpha^2+\beta^4)z^2(z^2-1)}{\left(\alpha^2+\beta^2(z^2-1)-(z^2-1)^2\right)^3}-\frac{2(3\beta^2+2)z^2-2\beta^2-4}{\left(\alpha^2+\beta^2(z^2-1)-(z^2-1)^2\right)^2}\right].$$

[1] We will use the results of Olver (1974, Chap. 11) extensively in this section, and therefore we will refer to this reference simply as "Olver" to avoid undue repetition.

In the second of (4.1) we take the branches to have their principal values when $z\in(1,z_2)$ and $\hat{\zeta}\in(0,\infty)$, and to be continuous elsewhere. We see that this transformation is a continuous 1-1 map of the variables z and $\hat{\zeta}$ from the domains $\hat{Z}$ to $\underline{\hat{Z}}$ (see Figs. 6a, b). We now record the asymptotic behavior of $z(\hat{\zeta})$ as $|\hat{\zeta}|\to\infty$ in $\underline{\hat{Z}}$. Firstly, from (4.1), we find that as $\hat{\zeta}\to\infty$ in the region between the curves DC and $D'C'$ of Fig. 6b, $z(\hat{\zeta})\to 1$ such that

$$z\sim 1+a_2(\alpha)\exp\left(-\frac{4}{3\alpha}\hat{\zeta}^{3/2}\right), \qquad (\alpha>0), \tag{4.4}$$

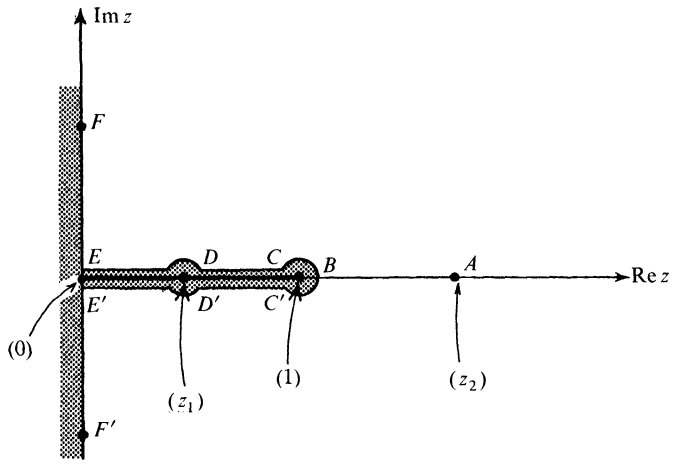


FIG. 6a. *Domain* $\hat{Z}$.

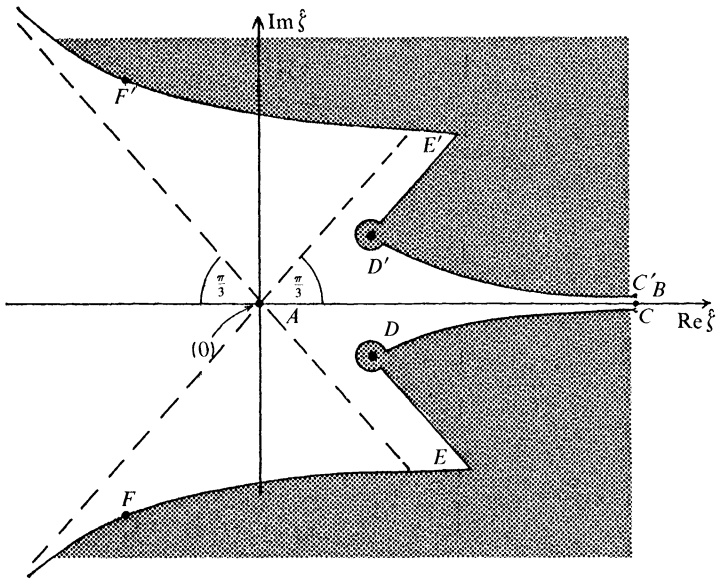


FIG. 6b. *Domain* $\underline{\hat{Z}}$.

where

$$a_2(\alpha)=2\left(\frac{z_2-1}{z_1+1}\right)e^{\rho_2(\alpha)},$$

with

$$\rho_2(\alpha)=\frac{2}{\alpha}\int_1^{z_2}\frac{\left[\left(t^2-z_1^2\right)\left(z_2^2-t^2\right)\right]^{1/2}-\alpha}{t^2-1}\,dt,$$

(cf. (2.10)). Also from (4.1) we find in a similar way to deriving (3.8) that $z\to\infty$ as $\hat{\xi}\to-\infty$, such that

$$z=\frac{2}{3}(-\hat{\xi})^{3/2}+\frac{\alpha\pi}{2}+\sigma+O\left(|\hat{\xi}|^{-3/2}\right), \tag{4.5}$$

where σ is given by (3.7b). From (4.3), (4.4) and (4.5) we see that

$$\hat{\psi}(\hat{\xi})=O(\hat{\xi}^{-2}) \tag{4.6}$$

when $\hat{\xi}\to\infty$ in both cases.

We now apply Theorem 9.1 of Olver to (4.2), with u replaced by γ, to obtain the following solutions holomorphic in $\hat{Z}$:

(4.7)

$$\hat{W}_{2n+1,j}(\gamma,\hat{\xi})=\mathrm{Ai}_j(\gamma^{2/3}\hat{\xi})\sum_{s=0}^{n}\frac{\hat{A}_s(\hat{\xi})}{\gamma^{2s}}+\frac{\mathrm{Ai}_j'(\gamma^{2/3}\hat{\xi})}{\gamma^{4/3}}\sum_{s=0}^{n-1}\frac{\hat{B}_s(\hat{\xi})}{\gamma^{2s}}+\hat{\varepsilon}_{2n+1,j}(\gamma,\hat{\xi}),$$

where n is a nonnegative integer, and

$$\mathrm{Ai}_j(z)=\mathrm{Ai}(ze^{-2\pi ij/3}),\qquad (j=0,\pm1).$$

Because of (4.6) the variations in the error bounds ((9.03) of Olver) converge in $\hat{Z}$. The reference points α_j associated with $\hat{W}_{2n+1,j}$ are taken to be $+\infty$, $\infty e^{2\pi i/3}$, $\infty e^{-2\pi i/3}$ for $j=0,1,-1$ respectively. With these choices of reference points we find from the known behavior of the Airy function at infinity that the solution $\hat{W}_{2n+1,j}$ is recessive as $\hat{\xi}\to\alpha_j$ $(j=0,\pm1)$. Therefore there exist constants $\hat{C}_{2n+1,j}(j=0,\pm1)$ such that

$$P{s}_\nu^{-\mu}(z,\gamma^2)=\hat{C}_{2n+1,0}\left(\frac{\hat{\xi}}{\alpha^2+\beta^2(z^2-1)-(z^2-1)^2}\right)^{1/4}\hat{W}_{2n+1,0}(\gamma,\hat{\xi}), \tag{4.8}$$

$$S_\nu^{\mu(3)}(z,\gamma^2)=\hat{C}_{2n+1,-1}\left(\frac{\hat{\xi}}{\alpha^2+\beta^2(z^2-1)-(z^2-1)^2}\right)^{1/4}\hat{W}_{2n+1,-1}(\gamma,\hat{\xi}), \tag{4.9}$$

$$S_\nu^{\mu(4)}(z,\gamma^2)=\hat{C}_{2n+1,1}\left(\frac{\hat{\xi}}{\alpha^2+\beta^2(z^2-1)-(z^2-1)^2}\right)^{1/4}\hat{W}_{2n+1,1}(\gamma,\hat{\xi}), \tag{4.10}$$

since, in each of these three equations, the functions on both sides of the equation are solutions of (1.1) that share the same recessive property at $\hat{\xi}=\alpha_0$, α_{-1} and α_1 respectively.

If we set the integration constants in $\hat{W}_{2n+1,0}$ so that

$$\hat{A}_s(\infty)=0,\qquad (s=1,2,\cdots) \tag{4.11}$$

and allow z to tend to 1 (and correspondingly $\hat{\xi} \to \infty$) in (4.8) we find, on employing (3.13), (4.4), (4.7) with (1.07) of Olver, that

$$\hat{C}_{2n+1,0} = A_\nu^{-\mu}(\gamma^2) \frac{2(\mu\pi)^{1/2}}{\gamma^{1/3}\Gamma(1+\mu)} \left(\frac{a_2(\alpha)}{2}\right)^{\mu/2} \left(1 - \frac{1}{\gamma} \sum_{s=0}^{n-1} \frac{\beta_s^+}{\gamma^{2s}}\right)^{-1}, \tag{4.12}$$

where

$$\beta_s^+ = \lim_{\hat{\xi} \to \infty} \{\hat{\xi}^{1/2} \hat{B}_s(\hat{\xi})\}, \qquad (s = 0, 1, 2, \cdots).$$

Similarly on letting $\hat{\xi} \to \alpha_{-1}$ in (4.9) and $\hat{\xi} \to \alpha_1$ in (4.10), and using (3.16), (4.5), (4.7) with (1.07) of Olver, we have

$$\hat{C}_{2n+1,-1} = 2\pi^{1/2}\gamma^{-5/6} e^{i(\chi - \pi/3)} \left(1 - \sum_{s=0}^{n-1} \frac{\beta_s^-}{\gamma^{2s+1}}\right)^{-1}, \tag{4.13}$$

and

$$\hat{C}_{2n+1,1} = 2\pi^{1/2}\gamma^{-5/6} e^{-i(\chi - \pi/3)} \left(1 - \sum_{s=0}^{n-1} \frac{\beta_s^-}{\gamma^{2s+1}}\right)^{-1}, \tag{4.14}$$

where

$$\beta_s^- = \lim_{\hat{\xi} \to \infty} \{(-\hat{\xi})^{1/2} \hat{B}_s(\hat{\xi})\}, \qquad (s = 0, 1, 2, \cdots).$$

We have set the integration constants in $\hat{W}_{2n+1,-1}$ and $\hat{W}_{2n+1,1}$ so that

$$\hat{A}_s(-\infty) = 0, \qquad (s = 1, 2, \cdots).$$

5. The characteristic exponent. There are four parameters associated with the spheroidal wave equation, namely λ, μ, γ and the characteristic exponent ν. There exists a transcendental relation among the four parameters; this relation is usually expressed in two ways: ν considered as a function of λ, μ, γ, or $\lambda \equiv \lambda_\nu^\mu(\gamma^2)$ considered as a function of ν, μ, γ. In the special case $\gamma = 0$, the spheroidal wave equation degenerates to the associated Legendre equation and the transcendental relationship is known to be

$$\lambda_\nu^\mu(0) = (\nu + 2p)(\nu + 2p + 1),$$

or, alternatively

$$\nu + 2p = \tfrac{1}{2}\{(4\lambda + 1)^{1/2} - 1\}, \qquad (p \in \mathbb{Z}).$$

For general values of the parameters, however, there is no explicit relationship in the literature. But we can obtain an asymptotic relationship between ν and λ, μ, γ, for large γ and λ, using the results of §3. This is of particular importance in problems of high-frequency (γ large) scattering by prolate spheroids; as an illustration of the problem the reader is referred to a paper by Sleeman (1969).

We start by considering the well-known analytic continuation formulae (Meixner and Schäfke (1954, p. 293))

$$S_\nu^{\mu(3)}(ze^{\pi i},\gamma)=e^{-\nu\pi i}S_\nu^{\mu(4)}(z,\gamma), \tag{5.1}$$

$$S_\nu^{\mu(4)}(ze^{\pi i},\gamma)=e^{\nu\pi i}S_\nu^{\mu(3)}(z,\gamma)+2i\sin(\nu\pi)S_\nu^{\mu(4)}(z,\gamma), \tag{5.2}$$

where $S_\nu^{\mu(3,4)}(ze^{\pi i})$ denotes the value of $S_\nu^{\mu(3,4)}(z)$ after describing a negative half-circuit about infinity. We now replace the S-functions in (5.1) and (5.2) with their uniform approximations of §3 to obtain an asymptotic expression for $\sin(\nu\pi)$; as will be seen shortly it is convenient to set $z=0-i0$ in (5.1) and (5.2). Now firstly from (3.15), (3.18), (3.20) and (3.21) we have

$$S_\nu^{\mu(3,4)}(0-i0,\gamma)=C_{2n+1}e^{\pm i\chi}\left\{\mathscr{H}_{2n+1}(\gamma,\alpha,\zeta_0)+\zeta_0^{-1/2}\varepsilon_{2n+1}^{\mp}(\gamma,\alpha,\zeta_0)\right\}, \tag{5.3}$$

and

$$S_\nu^{\mu(3,4)}(0+i0,\gamma)=C_{2n+1}e^{\pm i\chi}\left\{\mathscr{H}_{2n+1}(\gamma,\alpha,\zeta_0e^{2\pi i})-\zeta_0^{-1/2}\varepsilon_{2n+1}^{\mp}(\gamma,\alpha,\zeta_0e^{2\pi i})\right\}, \tag{5.4}$$

where

$$C_{2n+1}\equiv C_{2n+1}^{-}e^{-i\chi}\left(\frac{\zeta_0-\alpha^2}{1+\beta^2-\alpha^2}\right)^{1/4}=C_{2n+1}^{+}e^{i\chi}\left(\frac{\zeta_0-\alpha^2}{1+\beta^2-\alpha^2}\right)^{1/4},$$

and

$$\mathscr{H}_{2n+1}(\gamma,\alpha,\zeta)\equiv H_\mu^{(1)}(\gamma\zeta^{1/2})\sum_{s=0}^{n}\frac{A_s(\alpha,\zeta)}{\gamma^{2s}}+\frac{\zeta^{1/2}}{\gamma}H_\mu^{(1)\prime}(\gamma\zeta^{1/2})\sum_{s=0}^{n-1}\frac{B_s(\alpha,\zeta)}{\gamma^{2s}},\qquad(\mu=\gamma\alpha). \tag{5.5}$$

Note that in (5.4) we have used the identity

$$\zeta(0+i0)=\zeta(0-i0)e^{2\pi i}\equiv\zeta_0e^{2\pi i};$$

in other words a negative half-circuit about infinity from $0-i0$ to $0+i0$ in the z-plane corresponds to a positive circuit from ζ_0 to $\zeta_0e^{2\pi i}$ in the ζ-plane. Indeed, it is this particular property that led us to set $z=0-i0$ in (5.1) and (5.2).

In order that the error bounds (3.11b) can be applied to ε^+ and ε^- in (5.3) and (5.4) we see from Figs. 4b and 5b that we must impose the restriction on α and β that ζ_0 and $\zeta_0e^{2\pi i}$ must lie in $\underline{\Delta}^+\cap\underline{\Delta}^-$. In §3 we saw that this is equivalent to the condition (3.3b); from (3.1) one can show that (3.3b) is equivalent to the condition

$$\int_{z_1}^{z_2}\frac{\left[(t^2-z_1^2)(z_2^2-t^2)\right]^{1/2}}{t^2-1}\,dt>\frac{1}{\gamma}I(\mu), \tag{5.6a}$$

where

$$I(\mu)=\int_{\mu^2(\tau^2-1)}^{\hat{X}_\mu^2(\tau^2-1)}\frac{(\mu^2+t)^{1/2}}{2t}\,dt. \tag{5.6b}$$

From the definition of $\hat{X}_\mu$ (B&D, §5) we have $\hat{X}_\mu=\mu+O(\mu^{-1/3})$ as $\mu\to\infty$, and $\hat{X}_\mu>\mu$ for $0\leqq\mu<\infty$. It follows then that $I(\mu)$ is positive and bounded for $0\leqq\mu<\infty$.

Now if we choose the integration constants in (3.10) so that

$$A_s(\alpha,\zeta_0)=0, \qquad (s=1,2,\cdots,n),$$

and use the following relation for Hankel functions (with $^-$ denoting complex conjugate)

$$H_\mu^{(1)}(ze^{\pi i})=-e^{-\mu\pi i}H_\mu^{(2)}(z)=-e^{-\mu\pi i}\overline{H}_\mu^{(1)}(\bar{z}), \qquad (\mu \text{ real}),$$

(see, for example, Olver (1974, pp. 238, 239)), we obtain from (5.5) the following equations

$$\mathscr{H}_{2n+1}(\gamma,\alpha,\zeta_0)=H_\mu^{(1)}(\gamma\zeta_0^{1/2})+\frac{\zeta_0^{1/2}}{\gamma}H_\mu^{(1)\prime}(\gamma\zeta_0^{1/2})\sum_{s=0}^{n-1}\frac{B_s(\alpha,\zeta_0)}{\gamma^{2s}}, \tag{5.7}$$

and

$$\mathscr{H}_{2n+1}(\gamma,\alpha,\zeta_0e^{2\pi i})=-e^{-\mu\pi i}\overline{\mathscr{H}}_{2n+1}(\gamma,\alpha,\zeta_0). \tag{5.8}$$

From (3.19), (5.1), (5.3), (5.4), (5.8) and (A1) we derive

$$2\,\mathrm{Re}\left[e^{-i(\gamma\sigma-3\pi/4)}\left\{\mathscr{H}_{2n+1}(\gamma,\alpha,\zeta_0)+\zeta_0^{-1/2}\varepsilon_{2n+1}^{+}(\gamma,\alpha,\zeta_0)\right\}\right]=0, \tag{5.9}$$

and therefore the *exact* relation for each n

$$\arg\left\{\mathscr{H}_{2n+1}(\gamma,\alpha,\zeta_0)+\zeta_0^{-1/2}\varepsilon_{2n+1}^{+}(\gamma,\alpha,\zeta_0)\right\}=\gamma\sigma-\frac{\pi}{4}+p_n\pi, \tag{5.10}$$

where p_n is some integer which we determine shortly.

Next from (3.19), (5.2), (5.3), (5.4), (5.8) and (A1) we derive

$$\begin{aligned}\sin(\nu\pi)=e^{i(\gamma\sigma-\pi/4)}&\left\{\mathscr{H}_{2n+1}(\gamma,\alpha,\zeta_0)+\zeta_0^{-1/2}\varepsilon_{2n+1}^{+}(\gamma,\alpha,\zeta_0)\right\}^{-1}\\ &\times\mathrm{Re}\left[e^{i(\mu\pi+\gamma\sigma-3\pi/4)}\left\{\mathscr{H}_{2n+1}(\gamma,\alpha,\zeta_0)+\zeta_0^{-1/2}\varepsilon_{2n+1}^{-}(\gamma,\alpha,\zeta_0)\right\}\right],\end{aligned} \tag{5.11}$$

and on employing the relation (5.10) this reduces to the following expression for $\sin(\nu\pi)$:

$$\begin{aligned}\sin(\nu\pi)=e^{-p_n\pi i}&\left|\mathscr{H}_{2n+1}(\gamma,\alpha,\zeta_0)+\zeta_0^{-1/2}\varepsilon_{2n+1}^{+}(\gamma,\alpha,\zeta_0)\right|^{-1}\\ &\times\mathrm{Re}\left[e^{i(\mu\pi+\gamma\sigma-3\pi/4)}\left\{\mathscr{H}_{2n+1}(\gamma,\alpha,\zeta_0)+\zeta_0^{-1/2}\varepsilon_{2n+1}^{-}(\gamma,\alpha,\zeta_0)\right\}\right].\end{aligned} \tag{5.12}$$

This is the first important result of this section. We will show shortly that $p_n=0$, provided γ is sufficiently large for each value of n. Equation (5.12) gives then an asymptotic expansion for $\sin(\nu\pi)$ for large γ. With the restrictions (1.6) and (5.6a) applying the error terms ε^+ and ε^- in this equation can be bounded by (3.11b).

Note that the Hankel functions in (5.12) are of the form

$$H_\mu^{(1)}(\mu\omega)$$

where ω is real, positive and greater than 1, and therefore they are of oscillatory nature as $\mu\to\infty$. We shall now replace these functions in (5.12) with their "Debye"-type approximations to obtain a more simple approximation for $\sin(\nu\pi)$. The price we pay for this is that the expression we obtain contains more error terms; bounds are given.

For simplicity we set $n=1$ in (5.12). The Debye approximations for the Hankel functions can be expressed as

(5.13)

$$H_{\gamma\alpha}^{(1)}(\gamma\zeta_0^{1/2})=\left(\frac{2}{\gamma\pi}\right)^{1/2}(\zeta_0-\alpha^2)^{-1/4}e^{i(\gamma\sigma-\pi/4)}\left\{1-\frac{i}{24\gamma}(3\zeta_0+2\alpha^2)(\zeta_0-\alpha^2)^{-3/2}+\delta_0\right\},$$

and

$$H_{\gamma\alpha}^{(1)\prime}(\gamma\zeta_0^{1/2})=i\left(\frac{2}{\gamma\pi}\right)^{1/2}\zeta_0^{-1/2}(\zeta_0-\alpha^2)^{1/4}e^{i(\gamma\sigma-\pi/4)}\{1+\delta_0'\}, \tag{5.14}$$

where δ_0 and δ_0' are constants which satisfy the following bounds for $\alpha^2<\zeta_0$ (Olver (1974, p. 377)):

$$|\delta_0|\leqq\frac{2}{\gamma^2}\mathscr{V}_{0,q}(U_2)\exp\left\{\frac{2}{\gamma}\mathscr{V}_{0,q}(U_1)\right\}, \tag{5.15}$$

(5.16)

$$|\delta_0'|\leqq\frac{1}{2\gamma}\zeta_0(\zeta_0-\alpha^2)^{-3/2}+\left(1+\frac{\zeta_0^2}{4\gamma^2}(\zeta_0-\alpha^2)^{-3}\right)^{1/2}\frac{2}{\gamma}\mathscr{V}_{0,q}(U_1)\exp\left\{\frac{2}{\gamma}\mathscr{V}_{0,q}(U_1)\right\},$$

with

$$q=-i(\zeta_0-\alpha^2)^{-1/2},\qquad U_1=\frac{(3q-5\alpha^2q^3)}{24},$$

and

$$U_2=\frac{81q^2-462\alpha^2q^4+385\alpha^4q^6}{1152}.$$

Now from (5.7), (5.13), (5.14) and

$$B_0(\alpha,\zeta_0)=-(\zeta_0-\alpha^2)^{-1/2}\int_{\alpha^2}^{\zeta_0}(t-\alpha^2)^{-1/2}\psi(\alpha,t)\,dt$$

(see (2.9a) of B&D) we have

$$\begin{aligned}\mathscr{H}_3(\gamma,\alpha,\zeta_0)+\zeta_0^{-1/2}\varepsilon_3^{+}(\gamma,\alpha,\zeta_0)\\ =\left(\frac{2}{\gamma\pi}\right)^{1/2}(\zeta_0-\alpha^2)^{-1/4}e^{i(\gamma\sigma-\pi/4)}\left(1-\frac{i}{\gamma}b(\alpha)\right)(1+\eta^{\pm}(\gamma,\alpha)),\end{aligned} \tag{5.17}$$

where

$$b(\alpha)=\int_{\alpha^2}^{\zeta_0}(t-\alpha^2)^{-1/2}\psi(\alpha,t)\,dt+\frac{1}{24}(3\zeta_0+2\alpha^2)(\zeta_0-\alpha^2)^{-3/2}, \tag{5.18}$$

$$\begin{aligned}\tilde{\varepsilon}^{\pm}(\gamma,\alpha)=\delta_0-\frac{i}{\gamma}\delta_0'\int_{\alpha^2}^{\zeta_0}(t-\alpha^2)^{-1/2}\psi(\alpha,t)\,dt\\ +\left(\frac{\pi\gamma}{2\zeta_0}\right)^{1/2}(\zeta_0-\alpha^2)^{1/4}e^{i(\pi/4-\gamma\sigma)}\varepsilon_3^{\pm}(\gamma,\alpha,\zeta_0),\end{aligned} \tag{5.19}$$

and

$$\eta^{\pm}(\gamma,\alpha)=\frac{\tilde{\varepsilon}^{\pm}(\gamma,\alpha)}{1-i\gamma^{-1}b(\alpha)}. \tag{5.20}$$

We see from (5.7), (5.13) and (5.14) that $p_n=0$ in (5.10) provided, for each value of n, γ is sufficiently large.

It follows then from (5.12) and (5.17) that ν satisfies

$$\sin(\nu\pi)=\left|\frac{1+\eta^{-}}{1+\eta^{+}}\right|\sin\left(\mu\pi+2\gamma\sigma-\tan^{-1}\frac{b}{\gamma}-\frac{\pi}{2}+\arg(1+\eta^{-})\right). \tag{5.21}$$

This is the second important result of this section. Note that from (5.20) and Jordan's inequality

$$|\varphi|\leqq\frac{\pi}{2}|\tan(\varphi)|, \qquad \left(0\leqq\varphi<\frac{\pi}{2}\right)$$

we obtain the following bounds for sufficiently large γ:

$$|\eta^{\pm}|\leqq|\tilde{\varepsilon}^{\pm}|, \tag{5.22}$$

$$|\arg(1+\eta^{-})|\leqq\frac{\pi}{2}|\tan(\arg(1+\eta^{-}))|\leqq\frac{\pi}{2}|\eta^{-}|. \tag{5.23}$$

Equation (5.21) then provides an asymptotic relationship between ν, λ, μ and γ, for large γ and λ, with the restrictions (1.6) and (5.6a) applying. All the error terms in (5.21) have explicit bounds under these restrictions and they are $O(1/\gamma^2)$ for large γ.

From (1.2), (1.5), (3.7b) and (5.21) we have in terms of the original parameters the relation

$$\nu_r=\theta+\varepsilon, \tag{5.24}$$

where ν_r denotes the range of numbers $-\nu-1+2r$ and $\nu+2r$ ($r=0,\pm1,\pm2,\cdots$) and where

$$\theta\equiv\mu+\frac{2}{\pi}\int_0^{z_1}\frac{\left[\gamma^2(1-t^2)^2+\lambda(1-t^2)-\mu^2\right]^{1/2}}{1-t^2}\,dt-\frac{1}{\pi}\tan^{-1}\frac{b}{\gamma}-\frac{1}{2}, \tag{5.25}$$

$$z_1=\frac{1}{\gamma}\left[\gamma^2-\frac{1}{2}\left((\lambda^2+4\mu^2\gamma^2)^{1/2}-\lambda\right)\right]^{1/2}, \tag{5.26}$$

and ε is given implicitly by the equation

$$\left|\frac{1+\eta^{-}}{1+\eta^{+}}\right|\sin(\theta\pi+\arg(1+\eta^{-}))=\sin(\theta\pi+\varepsilon\pi). \tag{5.27}$$

An upper bound for $|\sin(\varepsilon\pi)|$ is readily derived from (5.27), but we will not record details here. However, from this bound we find that

$$\sin(\varepsilon\pi)=O\left(\min\left\{\gamma^{-2}|\cos(\theta\pi)|^{-1},\gamma^{-1}\right\}\right). \tag{5.28}$$

An interesting observation from (5.27), (5.28) is that if δ is a fixed arbitrary number such that

$$|\cos(\theta\pi)|\geqq\delta>0, \tag{5.29}$$

then, for sufficiently large γ, ε is real and $O(1/\gamma^2)$, and therefore from (5.24) it is seen that ν is real in the same circumstances.

Finally, let us set $\mu=0$ in (5.24) and (5.25). We obtain in this particular case the relation

$$\nu_r \sim \frac{2}{\pi}(\gamma^2+\lambda)^{1/2} E\left(\frac{\gamma}{(\gamma^2+\lambda)^{1/2}}\right)-\frac{1}{2},$$

where E is the complete elliptic integral of the second kind (see Gradshteyn and Ryzhik (1980, p. 905)). This approximation is the same as the eigenvalue equation (3.24) of Miles (1975). (In Miles' notation $\gamma^2 \Leftrightarrow c^2$, $\nu \Leftrightarrow n$, $\lambda+\gamma^2 \Leftrightarrow \lambda$ and $\lambda/\gamma^2 \Leftrightarrow -\beta^2$.)

6. Uniform approximations for the spheroidal functions $Qs_\nu^{-\mu}(x,\gamma^2)$ and $Qs_\nu^\mu(z,\gamma^2)$. In this section we derive asymptotic expansions for large γ for the spheroidal functions Qs, Qs. The results will be uniformly valid in the real case for $0 \leqq x < 1$, and in the complex case for $z \in \Delta^+ \cap \Delta^-$ (see Figs. 4a, 5a). They will also be uniformly valid for all values of ν, λ, μ, γ satisfying the restrictions (1.6), (5.6a) and (5.29). (The last restriction will be explained shortly.)

We first turn to the problem of deriving a uniform approximation for the spheroidal function Qs. Unfortunately this function is not recessive anywhere in the complex plane and therefore we cannot identify directly with the functions given by (3.10). We overcome this difficulty by using the fact that

$$Qs_\nu^\mu(z,\gamma^2) \propto \left\{e^{\nu\pi i} S_\nu^{\mu(3)}(z,\gamma) - e^{-\nu\pi i} S_\nu^{\mu(4)}(z,\gamma)\right\}, \tag{6.1}$$

and then replacing the RHS of (6.1) with the uniform approximations of §3 for the S-functions. However, a difficulty arises in this method if ν is close to $\frac{1}{2}$ (mod 1), for then severe cancellations take place on the right-hand side of (6.1) in the domain $\Delta^+ \cap \Delta^-$ where both S-functions are dominant. (From (5.21) and (5.25) we see that $\nu \equiv \frac{1}{2}$ (mod 1) implies that $\cos(\theta\pi)$ is $O(1/\gamma^2)$ for large γ.) As Arscott (1964) remarks, the case when $\nu \equiv \frac{1}{2}$ (mod 1) is exceptional in a very large part of the theory of spheroidal functions; it appears that this case has not been treated in much detail in the literature. The results in this section are valid for ν bounded away from $\frac{1}{2}$ (mod 1) for large γ.

It follows now from (3.10), (3.15), (3.18), (3.20), (3.21) and (6.1) that there exists a constant C_{2n+1}^* such that

$$Qs_\nu^\mu(z,\gamma^2) = C_{2n+1}^* \left(\frac{\alpha^2-\zeta}{\alpha^2+\beta^2(z^2-1)-(z^2-1)^2}\right)^{1/4} \times \left\{H_\mu^{(1)}(\gamma\zeta^{1/2}) \sum_{s=0}^{n} \frac{A_s(\alpha,\zeta)}{\gamma^{2s}} + \frac{\zeta^{1/2}}{\gamma} H_\mu^{(1)\prime}(\gamma\zeta^{1/2}) \sum_{s=0}^{n-1} \frac{B_s(\alpha,\zeta)}{\gamma^{2s}} + \zeta^{-1/2}\varepsilon_{2n+1}^*(\gamma,\alpha,\zeta)\right\}, \tag{6.2}$$

where

$$\varepsilon_{2n+1}^*(\gamma,\alpha,\zeta) = \frac{1}{2i}\operatorname{cosec}(\nu\pi+\chi)\left\{e^{i(\nu\pi+\chi)}\varepsilon_{2n+1}^+(\gamma,\alpha,\zeta) - e^{-i(\nu\pi+\chi)}\varepsilon_{2n+1}^-(\gamma,\alpha,\zeta)\right\}. \tag{6.3}$$

From (3.11b) and (6.3) we have the following bounds for ε^*:

(6.4)
$$\frac{|\varepsilon^*_{2n+1}(\gamma,\alpha,\zeta)|}{\zeta^{1/2}M^{(2)}_{\gamma\alpha}(\gamma\zeta^{1/2})},\ \frac{|\partial\varepsilon^*_{2n+1}(\gamma,\alpha,\zeta)/\partial\zeta|}{\frac{1}{2}\gamma N^{(2)}_{\gamma\alpha}(\gamma\zeta^{1/2})+\frac{1}{2}\zeta^{-1/2}M^{(2)}_{\gamma\alpha}(\gamma\zeta^{1/2})}$$
$$\leqq \frac{\kappa E^{(1)}_{\gamma\alpha}(\gamma\zeta^{1/2})^{-1}}{2\gamma^{2n+1}|\sin(\nu\pi+\chi)|}$$
$$\times\left[\mathscr{V}_{\zeta,-\infty+i0}\left\{(\xi-\alpha^2)^{1/2}B_n(\xi)\right\}\exp\left(\frac{\kappa}{\gamma}\mathscr{V}_{\zeta,-\infty+i0}\left\{(\xi-\alpha^2)^{1/2}B_0(\xi)\right\}\right)\right.$$
$$\left.+\mathscr{V}_{\zeta,-\infty-i0}\left\{(\xi-\alpha^2)^{1/2}B_n(\xi)\right\}\exp\left(\frac{\kappa}{\gamma}\mathscr{V}_{\zeta,-\infty-i0}\left\{(\xi-\alpha^2)^{1/2}B_0(\xi)\right\}\right)\right],$$

when $\zeta\in\underline{\Delta}^+\cap\underline{\Delta}^-$ (see Figs. 4b and 5b). In deriving (6.4) we have made use of the fact that ν is real—a consequence of the restriction (5.29).

Now from (5.24) and the definition (3.19) of χ we have

$$\sin(\nu\pi+\chi)=-\cos\left(\theta\pi+\frac{1}{2}\tan^{-1}\frac{b(\alpha)}{\gamma}+\frac{\varepsilon\pi}{2}\right), \tag{6.5}$$

and so (5.29) and (6.5) give the bound

$$\frac{1}{|\sin(\nu\pi+\chi)|}\leqq\frac{1}{\delta-|\sin((1/2)\tan^{-1}(b(\alpha)/\gamma)+\varepsilon\pi/2)|}. \tag{6.6}$$

The bounds (6.4) and (6.6) together establish the uniform validity of (6.2) for $z\in\Delta^+\cap\Delta^-$. The constant of proportionality C^*_{2n+1} could be determined by comparing both sides of (6.2) as $z\to 1$ and $\zeta(z)\to 0$. However, the expression for C^*_{2n+1} derived by this method would include a constant for which only an explicit upper bound could be given (cf. B&D, eq. (4.18)).

We therefore determine C^*_{2n+1} as follows. Firstly we use the relation (Meixner and Schäfke (1954, p. 287))

(6.7)
$$i\pi\frac{\Gamma(\nu+\mu+1)}{\Gamma(\nu-\mu+1)}\mathrm{P}s^{-\mu}_\nu(y,\gamma^2)=-e^{-\mu\pi i/2}Qs^\mu_\nu(y+i0,\gamma^2)+e^{-3\mu\pi i/2}Qs^\mu_\nu(y-i0,\gamma^2),$$
$$(-1<y<1),$$

with (2.14), (6.2) and (A6) to obtain the following:

$$i\pi\frac{\Gamma(\nu+\mu+1)}{\Gamma(\nu-\mu+1)}C_{2n+1,1}\left\{\mathscr{J}_{2n+1}(\gamma,\alpha,y)+y^{-1/2}\varepsilon_{2n+1,1}(\gamma,\alpha,y)\right\} \tag{6.8}$$
$$=2e^{-3\mu\pi i/2}C^*_{2n+1}\left\{\mathscr{J}_{2n+1}(\gamma,\alpha,y)+\mathrm{Re}\left[(y+i0)^{-1/2}\varepsilon^*_{2n+1}(\gamma,\alpha,y+i0)\right]\right\},$$

where

$$\mathscr{J}_{2n+1}(\gamma,\alpha,y)\equiv J_\mu(\gamma y^{1/2})\sum_{s=0}^{n}\frac{A_s(\alpha,y)}{\gamma^{2s}}+\frac{y^{1/2}}{\gamma}J'_\mu(\gamma\zeta^{1/2})\sum_{s=0}^{n-1}\frac{B_s(\alpha,y)}{\gamma^{2s}}. \tag{6.9}$$

We observe that the second term on the right-hand side of (6.8) must be recessive at $y=0$ since all other terms in this equation are recessive at $y=0$. From (6.3), (A2) and (A3) we therefore deduce that

(6.10)

$$\mathrm{Re}\big(\varepsilon^*_{2n+1}(y+i0)\big)=\int_0^y K(y,\xi)\Big[-\frac{1}{\gamma^{2n}}\big\{(\alpha^2-\xi)^{1/2}B_n(\alpha,\xi)\big\}'\xi^{1/2}J_\mu(\gamma\xi^{1/2})$$

$$+(\alpha^2-\xi)^{-1/2}\psi(\alpha,\xi)\mathrm{Re}\big(\varepsilon^*_{2n+1}(\xi+i0)\big)\Big]\,d\xi,$$

where

$$K(y,\xi)=\pi(\alpha^2-\xi)^{1/2}\big\{y^{1/2}Y_\mu(\gamma y^{1/2})\xi^{-1/2}J_\mu(\gamma\xi^{1/2})-y^{1/2}J_\mu(\gamma y^{1/2})\xi^{-1/2}Y_\mu(\gamma\xi^{1/2})\big\}.$$

This is precisely the integral equation satisfied by $\varepsilon_{2n+1,1}$ (B&D, eq. (3.3)) and therefore it follows that

$$\mathrm{Re}\big(\varepsilon^*_{2n+1}(\gamma,\alpha,y+i0)\big)=\varepsilon_{2n+1,1}(\gamma,\alpha,y).$$

From (6.8) we therefore have the *exact* relationship

$$C^*_{2n+1}=\frac{i\pi}{2}e^{3\mu\pi i/2}\frac{\Gamma(\nu+\mu+1)}{\Gamma(\nu-\mu+1)}C_{2n+1,1}. \tag{6.11}$$

We finally turn to the problem of establishing a uniform asymptotic expansion for $Qs_\nu^{-\mu}(x,\gamma^2)$ when $x\in[0,1)$. Given the relation (Meixner and Schäfke (1954, p. 287))

(6.12)

$$Qs_\nu^{-\mu}(x,\gamma^2)=\frac{\Gamma(\nu-\mu+1)}{2\Gamma(\nu+\mu+1)}\big(e^{-\mu\pi i/2}Qs_\nu^{\mu}(x+i0,\gamma^2)+e^{-3\mu\pi i/2}Q_\nu^{\mu}(x-i0,\gamma^2)\big)$$

one deduces from (6.2), (6.11) and (A6) that

$$Qs_\nu^{-\mu}(x,\gamma^2)=-\frac{\pi}{2}C_{2n+1,1}\left(\frac{\alpha^2-\zeta}{\alpha^2-\beta^2(1-x^2)-(1-x^2)^2}\right)^{1/4} \tag{6.13}$$

$$\times\Big[Y_\mu(\gamma\zeta^{1/2})\sum_{s=0}^{n}\frac{A_s(\alpha,\zeta)}{\gamma^{2s}}+\frac{\zeta^{1/2}}{\gamma}Y'_\mu(\gamma\zeta^{1/2})\sum_{s=0}^{n-1}\frac{B_s(\alpha,\zeta)}{\gamma^{2s}}$$

$$+\mathrm{Im}\big\{(\zeta+i0)^{-1/2}\varepsilon^*_{2n+1}(\gamma,\alpha,\zeta+i0)\big\}\Big].$$

7. Summary. For reference we now collect the principal results of this paper, re-expressed in terms of the original variables z (complex) or x (real), and the original parameters λ, μ, γ of the defining differential equation (1.1).

We introduce a transformed independent variable ζ that is an increasing function of x, and a continuous function of λ, μ, γ and z. When $z\equiv x$ is real with $0\leqq x<1$, ζ is

defined by equations (2.2) and (2.3). In these equations α, β, z_1 and z_2 are given by

$$\alpha=\frac{\mu}{\gamma}, \tag{7.1a}$$

$$\beta^2=\frac{\lambda}{\gamma^2}, \tag{7.1b}$$

$$z_1^2=1+\frac{\lambda}{2\gamma^2}-\frac{1}{2\gamma^2}\left(\lambda^2+4\mu^2\gamma^2\right)^{1/2}, \tag{7.2a}$$

$$z_2^2=1+\frac{\lambda}{2\gamma^2}+\frac{1}{2\gamma^2}\left(\lambda^2+4\mu^2\gamma^2\right)^{1/2}. \tag{7.2b}$$

When z is complex ζ is defined by (3.1).

Next, we introduce a function ψ which is given by (2.9). With this equation for ψ, functions $A_s(\zeta)$ and $B_s(\zeta)$ $(s=0,1,2,\cdots)$ are defined recursively by (2.9a, b) of B&D, in which the λ_s are chosen such that

$$A_s(0)=0, \qquad (s=1,2,3,\cdots). \tag{7.3}$$

The following expansions are uniformly valid for $\gamma>0$, $\mu\geqq 0$ and

$$0\leqq\frac{\mu^2}{\gamma^2}\leqq 1+A, \tag{7.4}$$

$$A'\leqq\frac{\lambda}{\gamma^2}\leqq A'', \tag{7.5}$$

where A, A' and A'' are arbitrary real constants such that $0\leqq A<A'\leqq A''<\infty$;

(7.6)

$$\mathrm{Ps}_\nu^{-\mu}(x,\gamma^2)=C_{2n+1,1}\left(\frac{\mu^2-\gamma^2\zeta}{\mu^2-\lambda(1-x^2)-\gamma^2(1-x^2)^2}\right)^{1/4}$$
$$\times\left[J_\mu(\gamma\zeta^{1/2})\sum_{s=0}^{n}\frac{A_s(\zeta)}{\gamma^{2s}}+\frac{\zeta^{1/2}}{\gamma}J_\mu'(\gamma\zeta^{1/2})\sum_{s=0}^{n-1}\frac{B_s(\zeta)}{\gamma^{2s}}+\zeta^{-1/2}\varepsilon_{2n+1,1}(\zeta)\right],$$

(7.7)

$$Ps_\nu^{-\mu}(z,\gamma^2)=\frac{1}{2}C_{2n+1,1}\left(\frac{\mu^2-\gamma^2\zeta}{\mu^2+\lambda(z^2-1)-\gamma^2(z^2-1)^2}\right)^{1/4}$$
$$\times\left[J_\mu(\gamma\zeta^{1/2})\sum_{s=0}^{n}\frac{A_s(\zeta)}{\gamma^{2s}}+\frac{\zeta^{1/2}}{\gamma}J_\mu'(\gamma\zeta^{1/2})\sum_{s=0}^{n-1}\frac{B_s(\zeta)}{\gamma^{2s}}+\zeta^{-1/2}\varepsilon_{2n+1}^{(0)}(\zeta)\right],$$

(7.8)

$$S_\nu^{\mu(3,4)}(z,\gamma)=C_{2n+1}^{\mp}\left(\frac{\mu^2-\gamma^2\zeta}{\mu^2+\lambda(z^2-1)-\gamma^2(z^2-1)^2}\right)^{1/4}$$
$$\times\left[H_\mu^{(1)}(\gamma\zeta^{1/2})\sum_{s=0}^{n}\frac{A_s(\zeta)}{\gamma^{2s}}+\frac{\zeta^{1/2}}{\gamma}H_\mu^{(1)\prime}(\gamma\zeta^{1/2})\sum_{s=0}^{n-1}\frac{B_s(\zeta)}{\gamma^{2s}}+\zeta^{-1/2}\varepsilon_{2n+1}^{\mp}(\zeta)\right].$$

The coefficient $C_{2n+1,1}$ is given by

$$C_{2n+1,1}=A_\nu^{-\mu}(\gamma^2)\left(\frac{1-z_1}{\mu^2}\right)^{\mu/2}(1+z_1)^{-\mu/2}\left(1+\frac{\mu}{\gamma^2}\sum_{s=0}^{n-1}\frac{B_s(0)}{\gamma^{2s}}\right)^{-1} \tag{7.9}$$
$$\times\exp\left\{\mu+\int_{z_1}^{1}\frac{\left[\mu^2-(1-t^2)-\gamma^2(1-t^2)^2\right]^{1/2}-\mu}{1-t^2}\,dt\right\}.$$

The coefficients $C_{2n+1}^{\pm}$ are given by

$$C_{2n+1}^{\pm}=ie^{\mp i\chi}\left(\frac{\pi}{2\gamma}\right)^{1/2}e^{\mu\pi i/2}\left(\sum_{s=0}^{n}\frac{A_s(-\infty)}{\gamma^{2s}}+\frac{1}{\gamma}\sum_{s=0}^{n-1}\frac{\tilde{B}_s(-\infty)}{\gamma^{2s}}\right)^{-1}, \tag{7.10}$$

in which

$$\chi=\frac{\mu\pi}{2}-\frac{\nu\pi}{2}+\int_0^{z_1}\frac{\left[\gamma^2(1-t^2)^2+\lambda(1-t^2)-\mu^2\right]^{1/2}}{1-t^2}\,dt-\frac{3\pi}{4}, \tag{7.11}$$

and

$$\tilde{B}_s(-\infty)=\lim_{\zeta\to-\infty}\left\{(-\zeta)^{1/2}B_s(\zeta)\right\}. \tag{7.12}$$

The error term $\varepsilon_{2n+1,1}$ in (7.6) is uniformly bounded by (3.8) of B&D for $0\leqq x<1$. The error term $\varepsilon_{2n+1}^{(0)}$ is uniformly bounded by (3.11a) for $z\in\Delta$ (see Fig. 1a). The error terms $\varepsilon_{2n+1}^{\pm}$ are uniformly bounded by (3.11b) for $z\in\Delta^{\pm}$ (see Figs. 4a and 5a).

A uniform relationship between λ, μ, γ and the characteristic exponent ν is given as follows. Firstly we define $\hat{X}_\mu$ to be the smallest root of

$$J_{\hat{\mu}}(x)+Y_{\hat{\mu}}(x)=0, \tag{7.13}$$

where $\hat{\mu}$ is the real (nonnegative) solution of

$$\mu=\hat{\mu}-c(\hat{\mu}/2)^{1/3} \tag{7.14}$$

in which c denotes the negative root of smallest absolute value of the equation $\mathrm{Ai}(x)=\mathrm{Bi}(x)$ ($c=-0.36605\cdots$). Then under the restrictions (7.4), (7.5), and in addition the restriction

$$\int_{z_1}^{z_2}\frac{\left[\mu^2+\lambda(t^2-1)-\gamma^2(t^2-1)^2\right]^{1/2}}{t^2-1}\,dt>\int_{\mu^2(\tau^2-1)}^{\hat{X}_\mu^2(\tau^2-1)}\frac{(\mu^2+t)^{1/2}}{2t}\,dt, \tag{7.15}$$

where τ is defined by (3.3a), we have the following:

$$\sin(\nu\pi)=\left|H_\mu^{(1)}(\gamma\zeta_0^{1/2})+\frac{\zeta_0^{1/2}}{\gamma}H_\mu^{(1)\prime}(\gamma\zeta_0^{1/2})\sum_{s=0}^{n-1}\frac{B_s(\zeta_0)}{\gamma^{2s}}+\zeta_0^{-1/2}\varepsilon_{2n+1}^{+}(\zeta_0)\right|^{-1} \tag{7.16}$$
$$\times\mathrm{Re}\left[\exp\left(i\left\{\mu\pi-\frac{3\pi}{4}+\int_0^{z_1}\frac{\left[\gamma^2(1-t^2)^2+\lambda(1-t^2)-\mu^2\right]^{1/2}}{1-t^2}\,dt\right\}\right)\right.$$
$$\left.\times\left\{H_\mu^{(1)}(\gamma\zeta_0^{1/2})+\frac{\zeta_0^{1/2}}{\gamma}H_\mu^{(1)\prime}(\gamma\zeta_0^{1/2})\sum_{s=0}^{n-1}\frac{B_s(\zeta_0)}{\gamma^{2s}}+\zeta_0^{-1/2}\varepsilon_{2n+1}^{-}(\zeta_0)\right\}\right].$$

In (7.16) ζ_0 denotes the value of $\zeta(x)$ at $x=0$, and $B_s(\zeta)$ $(s=0,1,2,\cdots,n-1)$ are given by (2.9a, b) of B&D, with λ_s chosen such that $A_s(\zeta_0)=0$ $(s=1,2,\cdots,n)$.

Under the restrictions (7.4), (7.5), (7.15), and in addition the restriction

$$\cos\left(\mu\pi+2\int_0^{z_1}\frac{\left[\gamma^2(1-t^2)^2+\lambda(1-t^2)-\mu^2\right]^{1/2}}{1-t^2}\,dt\right)\geqq\delta>0, \tag{7.17}$$

where δ is an arbitrary fixed number, we have the following uniform asymptotic expansions:

(7.18)

$$Qs_\nu^\mu(z,\gamma^2)=\frac{i\pi}{2}e^{3\mu\pi i/2}\frac{\Gamma(\nu+\mu+1)}{\Gamma(\nu-\mu+1)}C_{2n+1,1}\left(\frac{\mu^2-\gamma^2\zeta}{\mu^2+\lambda(z^2-1)-\gamma^2(z^2-1)^2}\right)^{1/4}$$
$$\times\left[H_\mu^{(1)}(\gamma\zeta^{1/2})\sum_{s=0}^{n}\frac{A_s(\zeta)}{\gamma^{2s}}+\frac{\zeta^{1/2}}{\gamma}H_\mu^{(1)\prime}(\gamma\zeta^{1/2})\sum_{s=0}^{n-1}\frac{B_s(\zeta)}{\gamma^{2s}}+\zeta^{-1/2}\varepsilon_{2n+1}^*(\zeta)\right],$$

(7.19)

$$Qs_\nu^{-\mu}(x,\gamma^2)=-\frac{\pi}{2}C_{2n+1,1}\left(\frac{\mu^2-\gamma^2\zeta}{\mu^2-\lambda(1-x^2)-\gamma^2(1-x^2)^2}\right)^{1/4}$$
$$\times\left[Y_\mu(\gamma\zeta^{1/2})\sum_{s=0}^{n}\frac{A_s(\zeta)}{\gamma^{2s}}+\frac{\zeta^{1/2}}{\gamma}Y_\mu'(\gamma\zeta^{1/2})\sum_{s=0}^{n-1}\frac{B_s(\zeta)}{\gamma^{2s}}\right.$$
$$\left.+\operatorname{Im}\left\{(\zeta+i0)^{-1/2}\varepsilon_{2n+1}^*(\zeta+i0)\right\}\right].$$

Equation (7.18) is uniformly valid for $z\in\Delta^+\cap\Delta^-$, and (7.19) is uniformly valid for $0\leqq x<1$. The error term ε_{2n+1}^* is uniformly bounded under the above-mentioned restrictions by (6.4).

Appendix. Analytic continuation of the error functions. We prove the following

$$\bar{\varepsilon}_{2n+1}^+(\gamma,\alpha,y\pm i0)=e^{\mu\pi i}\varepsilon_{2n+1}^-(\gamma,\alpha,y\mp i0),\qquad(0<y\leqq\zeta_0), \tag{A1}$$

where $^-$ denotes complex conjugate.

The error functions ε^+, ε^- are solutions of the integral equation

(A2)

$$\varepsilon_{2n+1}^\pm(\gamma,y,\zeta)=\int_{-\infty\pm i0}^{\zeta}K(\zeta,\xi)\left[-\frac{1}{\gamma^{2n}}\left\{(\xi-\alpha^2)^{1/2}B_n(\gamma,\xi)\right\}'\xi^{1/2}H_\mu^{(1)}(\gamma\xi^{1/2})\right.$$
$$\left.+(\xi-\alpha^2)^{-1/2}\psi(\alpha,\xi)\varepsilon_{2n+1}^\pm(\gamma,\alpha,\xi)\right]d\xi,$$

where

(A3)

$$K(\zeta,\xi)=\pi(\xi-\alpha^2)^{1/2}\left(\frac{\zeta}{\xi}\right)^{1/2}\left\{H_\mu^{(1)}(\gamma\zeta^{1/2})H_\mu^{(2)}(\gamma\xi^{1/2})-H_\mu^{(2)}(\gamma\zeta^{1/2})H_\mu^{(1)}(\gamma\xi^{1/2})\right\}.$$

If we set $\zeta = y \pm i0$ in (A2) and take conjugates of both sides, we have the following equation for $\bar{\varepsilon}^{+}$:

(A4)

$$\bar{\varepsilon}^{+}_{2n+1}(\gamma,\alpha,y\pm i0)=\int_{-\infty+i0}^{y\pm i0}\overline{K}(y\pm i0,\xi)\left[-\frac{1}{\gamma^{2n}}\left\{(\bar{\xi}-\alpha^2)^{1/2}B_n(\alpha,\bar{\xi})\right\}'\bar{\xi}^{1/2}\overline{H}^{(1)}_{\mu}(\gamma\xi^{1/2})\right.$$

$$\left.+(\bar{\xi}-\alpha^2)^{-1/2}\psi(\alpha,\bar{\xi})\bar{\varepsilon}^{+}_{2n+1}(\gamma,\alpha,\xi)\right]d\bar{\xi}.$$

If we make the change of variable in (A4) of

$$\omega=\bar{\xi}e^{2\pi i},\qquad (0\leqq\arg(\omega)<2\pi),$$

and use the identities

$$\overline{H}^{(1)}_{\mu}(\gamma\xi^{1/2})=-e^{\mu\pi i}H^{(1)}_{\mu}(\gamma\omega^{1/2}),$$

and $\overline{K}(y\pm i0,\xi)=K(y\mp i0,\omega)$, (see Olver (1974, pp. 238 and 239)), we find that

(A5) $$e^{-\mu\pi i}\bar{\varepsilon}^{+}_{2n+1}(\gamma,\alpha,y\pm i0)=\int_{-\infty-i0}^{y\pm i0}K(y\mp i0,\omega)$$

$$\times\left[-\frac{1}{\gamma^{2n}}\left\{(\omega-\alpha^2)^{1/2}B_n(\alpha,\omega)\right\}'\omega^{1/2}H^{(1)}_{\mu}(\gamma\omega^{1/2})\right.$$

$$\left.+(\omega-\alpha^2)^{-1/2}\psi(\alpha,\omega)e^{-\mu\pi i}\bar{\varepsilon}^{+}_{2n+1}(\gamma,\alpha,\bar{\omega}e^{2\pi i})\right]d\omega;$$

(A1) now follows from the observation that (A5) is precisely the integral equation (A2) satisfied by $\varepsilon^{-}_{2n+1}(\gamma,\alpha,y\mp i0)$.

Finally we find from (6.3) and (A1) that if ν is real (a sufficient condition is (5.29)) then the following analytic continuation formula for ε^{*} holds

(A6) $$\bar{\varepsilon}^{*}_{2n+1}(\gamma,\alpha,y+i0)=e^{\mu\pi i}\varepsilon^{*}_{2n+1}(\gamma,\alpha,y-i0).$$

Acknowledgment. The author is grateful to Dr. W. G. C. Boyd for suggesting the problem and for the invaluable help he gave during the writing of this paper.

REFERENCES

M. Abramowitz and I. A. Stegun (1965), *Handbook of Mathematical Functions*, 7th ed., Dover, New York.

F. M. Arscott (1964), *Periodic Differential Equations*, Pergamon Press, Oxford.

W. G. C. Boyd and T. M. Dunster (1986), *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*, this Journal, 17, pp. 422–450.

J. Des Cloizeaux and M. L. Mehta (1972), *Some asymptotic expressions for spheroidal wave functions and for the eigenvalues of differential and integral equations of which they are solutions*, J. Math. Phys., 13, pp. 1745–1754.

C. Flammer (1957), *Spheroidal Wave Functions*, Stanford Univ. Press, Stanford, CA.

I. S. Gradshteyn and I. M. Ryzhik (1980), *Tables of Integrals, Series and Products*, 4th ed., Academic Press, London.

J. Meixner and F. W. Schäfke (1954), *Mathieusche Funktionen und Sphäroid-funktionen*, Springer-Verlag, Berlin.

J. W. Miles (1975), *Asymptotic approximations for prolate spheroidal wave functions*, Stud. Appl. Math., 54, pp. 315–349.

H. J. W. MÜLLER (1963), *Asymptotic expansions of prolate spheroidal wave functions and their characteristic numbers*, J. Reine Angew. Math., 212, pp. 26–48.

F. W. J. OLVER (1974), *Asymptotics and Special Functions*, Academic Press, New York.

M. L. SINK AND B. C. EU (1983), *A uniform* WKB *approximation for spheroidal wave functions*, J. Chem. Phys., 78, pp. 4887–4895.

B. D. SLEEMAN (1969), *Integral representations associated with high-frequency non-symmetric scattering by prolate spheroids*, Quart. Appl. Math., 22, pp. 405–426.

D. SLEPIAN (1965), *Some asymptotic expansions for prolate spheroidal wave functions*, J. Math. and Phys., 44, pp. 99–140.

W. STREIFER (1968), *Uniform asymptotic expansions for prolate spheroidal wave functions*, J. Math. and Phys., 47, pp. 407–415.